SUPPLEMENTAL MATERIAL

Degrees of convergent evolution in rodent adaptations to arid environments

The supplemental material is structured in 7 parts, corresponding to sections of the main manuscript. All supplemental Figures, supplemental Tables, and supplemental Data are listed (some Supplemental Tables and data are too large and therefore provided separately).

Table of content

Appendix 1. Species sampling, data preparation and sequencing

Supplemental Methods: Species trapping, species choice, environment annotation

Sampling

Our sampling relied on animals trapped in the wild and animals belonging to colonies of wild species (*Acomys dimidiatus*, different *Mus* species, *Fukomys mechowii*). We also added a commercially available mouse strain, as a reference. Only one species was trapped on purpose for this study, others were trapped in the frame of other projects. In the latter case, kidneys were preserved for the purpose of this study, or they had been preserved as part of collections, and were kindly given to us for this project.

Sampling of kidneys from wild animals trapped for other projects.

Two rodent trapping campaigns occurred during the time of this project, during which our colleagues kindly accepted to preserve kidneys in RNA later for us. These are described below.

Established collections.

Several samples were also kindly given to us from collections of samples, established through years of research on rodents and field campaigns. This includes: *Mus Nannomys indutus* (Pascale Chevret LBBE, Lyon, France and Janice Britton-Davidian, ISEM, Montpellier, France, (Chevret et al. 2014)), *Apodemus mystacinus* (Petros Lymberakis, collection of the Natural History Museum of Crete under Presidential Decree 67/81, Greece, Michaux et al, 2005), *Mastomys kollmanspergeri* (Gauthier Dobigny, and Philippe Gauthier, IRD-CBGP, Montferrier-sur-Lez, France, (Savassi et al. 2021); *Mastomys natalensis* (Gauthier Dobigny and Madougou Garba, IRD-CBGP, Montferrier-sur-Lez, France, (Garba et al. 2014)), *Mus Nannomys* sp (Gauthier Dobigny, IRD-CBGP, Montferrier-sur-Lez, France). These *Mastomys* and *Nannomys* samples were part of the "CBGP - Small mammal Collection", https://doi.org/10.15454/WWNUPO.

Of note, two unidentified *Mus Nannomys* samples were kindly gifted by Gauthier Dobigny (IRD-CBGP) and sequenced as part of Bioproject (PRJEB54931), but were finally not used in our analyses as species identification was wrong (*Mus musculus*) or unclear (*Mus Nannomys sp*). Access and sharing of advantages had been agreed by the government of the Republic of Benin (file 608/DGEFC/DCPRN/PF-APA/SA).

Sampling for this study in the frame of other projects

Trapping in Senegal

Arvicanthis niloticus and *Mastomys erythroleucus* kidneys were kindly gifted by Laurent Granjon (IRD-CBGP, France).

Trapping sessions were conducted following a standardized protocol (see Granjon et al. 2019 for details) using locally made wire-mesh live traps (8.5 × 8.5 × 26.5 cm) and Sherman folding box traps (8 × 9 × 23 cm) baited with peanut butter and fresh onions. Within each site, traps were set each afternoon, checked for night captures the following morning and then re-baited for additional trap nights if necessary. Once caught, rodent identification was performed based on morphology in the field and further molecular diagnosis in the lab if necessary following Granjon & Duplantier (2009). They were euthanized by cervical dislocation and then weighed to the nearest 0.5 g, sexed, measured and dissected. None of the rodent species investigated here is officially protected, and every animal-related procedure was performed according to official ethical guidelines (Sikes and Animal Care and Use Committee of the American Society of Mammalogists 2016).

All protocols used here were conducted following official regulations (Centre de Biologie pour la Gestion des Populations (CBGP): Agrément pour l'utilisation d'animaux à des fins scientifiques D-34-169-003) of the relevant institutional committee (Regional Head of the Veterinary Service, Hérault, France). All transfer and conservation procedures were performed in accordance with current Senegalese and French legislation by the Senegalese "Direction Nationale des Eaux et Forêts, Chasses Et Conservation Des Sols", partner of the "CERISE" project which funded this trapping campaign, and the French "Direction Départementale de la Protection des Populations, Hérault" for sample importation authorization.

This work was supported by the CERISE project (funded by the Fond Français pour l'Environnement Mondial via the "Fondation pour la Recherche sur la Biodiversite": AAPSCEN-20B III).

Trapping in South Africa

Mastomys coucha, *Micaelamys namaquensis*, *Dendromus melanotis* kidneys were kindly sampled by Frédéric Delsuc, Lionel Hautier, and Nico Avenant.

The South African rodent specimens were sampled at Tussen die Riviere Nature Reserve (Free State, South Africa) in October 2017 as part of the ConvergeAnt project (ERC Consolidator Grant #683257). Sampling was conducted under permit number JM 1193/2017, issued by the Free State Department of Economic, Small Business Development, Tourism and Environmental Affairs (DESTEA) in Bloemfontein (Free State, South Africa). These samples have been sent to France under export permit JM 3007/2017, also issued by DESTEA. As these species are classified as Least Concern by the IUCN, and do not require CITES permits for international transport, the samples were transferred to France under import permits issued by the Direction régionale de l'environnement, de l'aménagement et du logement (DREAL) Occitanie in Toulouse (France).

Sampling of kidneys from wild animals trapped for this study

Trapping in Lyon, France (Pascale Chevret)

Trapping sessions were conducted on La Doua Campus (Villeurbanne) using INSA traps (5 x 5 x 16 cm). All procedures regarding animal handling complied with the approved guidelines by the American Society of Mammalogists (Sikes and Animal Care and Use Committee of the American Society of Mammalogists 2016). All *Apodemus* were killed by cervical dislocation in accordance with the European Parliament directive 2010/63/UE on the protection of animals used for scientific purposes.

Sampling of kidneys from lab-maintained animals killed for this study

Sampling in Lyon, France

Acomys dimidiatus is a well recognized arid species, and a colony is maintained in our lab for other purposes (Agreement for wild species delivered to Sophie Pantalacci; SPE-2014-001 #69-148). We took this opportunity to collect kidney samples. Animals were anesthetized with cevofluorane to enable intraperitoneal injection of a ketamine-xylasine mix, and then killed with pentobarbital administered intracardially.

In the same conditions, we also sampled a standard mouse lab strain (outbred, CD1, Charles River laboratory). These mice were killed by cervical dislocation.

These two species were maintained and sacrificed in strict accordance with the European guidelines 2010/63/UE.

Sampling in Montpellier, France

We also had the opportunity to select different *Mus* species from the “Conservatoire de la souris” (Montpellier, France), thanks to his director François Bonhomme. These are strains established from mice trapped at specific locations over the world, and maintained for generations. We initially selected: *Mus spretus* (STF, captured in Tunisia), *Mus macedonicus* (XBS, captured in Bulgaria), *Mus musculus domesticus* (WLA and DDO captured in south-west of France and Denmark respectively) and *Mus pahari* (PAH, captured in Thailande) (see Supplemental Table S2 for exact locations). Except for *Mus macedonicus*, which is well recognized as adapted to mediterranean climate, the status of other species (in particular *spretus* and *pahari*) may vary between populations. We measured their daily water consumption (normalized with body weight) and found that *macedonicus* indeed has the smallest water consumption (XBS:0.16ml/day/g), *spretus* was intermediate (STF:0.2ml/day/g), close enough from *pahari* strain (0.23ml/day/g), but also from one of the two *domesticus* strains (DDO: 0.24ml/day/g, WLA: 0.34ml/day/g). Facing this continuum, we finally decided to exclude *spretus* and *pahari* from our expression analysis to minimize *Mus* genus overrepresentation.

These mice were maintained and sacrificed by cervical dislocation in strict accordance with the European guidelines 2010/63/UE.

Sampling in České Budějovice, Czech republic

Fukomys mechowii samples were kindly given by Radim Sumbera (University of South Bohemia), who maintains a colony (accreditation number 22395/2014-MZE-17214). Dissections were realized by Lucie Plestilova.

Bioclimatic variables

We obtained the geographical distribution area of each species using GBIF (https://www.gbif.org/) data through the rgbif package (Chamberlain and Boettiger 2017). Then, for each species we extracted BIO17 values of its distribution area with the dismo package (Hijmans et al. 2010), which indicate precipitation values of the driest quarter, from the international database worldclim (https://www.worldclim.org/data/bioclim.html). Median values were calculated for each species after excluding samples from zoos, museums or laboratories. We considered a species as adapted to an arid environment if the median BIO17 is below 40 and mesic if the median BIO17 is over 40. For homogeneity, we also used the median of the species for the collected samples even if we have the variable for their location of capture. The biological status of the collected samples was similar whether taking the median of the species or the specific location of capture, except for *Mastomys natalensis* (Species-BIO17 is 46 and Sample-BIO17 is 0) which was only used for the sequence-based analyses due to the ambiguity of the status; and for Dendromus melanotis (Species-BIO17 is 83 and Sample-BIO17 is 39), which was annotated as a mesic species. The data is available in Supplemental Table S4.

Ancestral state reconstruction/ Independence of transitions

To obtain the ancestral state reconstruction, we opted for a semi-automated method, based on a broad annotation in many species, which we curated manually for certain nodes.

The annotation of the bioclimatic variables was intersected with a large scale tree of rodent species (2257 species, (Fabre et al. 2012), Supplemental Data S2). That phylogeny was inferred based on several genes (6 mitochondrial genes :12S rRNA, D-Loop, COX3, Cytb, NADH4, NADH1 and 5 nuclear genes : BRCA1, IRBP, GHR, RAG1, vWF). We used this phylogeny to ascertain the ancestral states with more power because it allows us to see potential reversions that would otherwise be masked if the corresponding species were not included. We also checked the congruence of the topology of our tree based on whole transcriptomes (shown in Fig 2). We found that the tree from Fabre et al. pruned to the same species were nearly identical except the relative branching order of *Mus spretus* and *Mus musculus* (cophylogenies were made with ape *(Paradis and Schliep 2019)* and phytool (Revell and Harmon 2022) R packages, Supplemental Fig. S2).

Each species on the tree was annotated with a discrete character, “arid” or “mesic”. Branches of null length were increased to a very small value (1e-08). We then reconstructed the ancestral states using the maximum likelihood estimation provided by the function ace (ape R package; (Paradis and Schliep 2019) with type = "discrete", model “equal rates”, using joint reconstruction. This provided us with the scaled likelihoods of each ancestral state. The plot for the full tree (using branch length and topology from (Fabre et al. 2012) is available in Supplemental Data S4 and a subtree with species for the total dataset is available in Supplemental Data S3. Because certain species have different names in GBIF and in the Fabre phylogeny, we modified them accordingly (see code available in supplemental code).

Based on this extensive automated annotation, we then reviewed individually the transitions and manually curated them, based on literature on rodent phylogeny and phylogeography (Table S5). After this manual curation, we obtained 22 transitions which are numbered on Supplemental Fig. S1, and individually discussed them in the text below, together with the help of annotations on the full tree presented in Supplemental Data S4. All discrepancies between the automatic reconstruction and the literature are pointed with red circles on Supplemental Fig. S1. These manual curations emphasized systematic biases in the automatic reconstruction. In several cases, major paleoclimate changes toward aridification drove the diversification of arid species, while mesic species were diversifying more slowly in their habitat, resulting in a marked desequilibrium in favor of arid species in the tree of current species. As a consequence, ancestral states tend to be biased toward arid states in the automatic reconstruction. In our manual curations, we therefore paid particular attention to the presence of basal mesic species or groups (pointed with green arrows on Supplemental Data S4) to make decisions on curation.

T1 : the transition is estimated to have occurred at the basis of the arid Dipodidae group (the *Jaculus jaculus* arid group), after the split with the mesic *Zapus* sister group (Zapodidae) and the mesic *Sicista* outgroup (Sminthidae). This interpretation is consistent with climate becoming drier after these splits after Oligocene-miocene transition in the region of origin for the Dipodidae group (Pisano et al. 2015).

T2/T3: The ancestral state reconstruction is ambiguous, likely due to phylogenetical uncertainty of the Fabre phylogeny with the Spalax species complex. According to He et al. 2020, spalacinae (*Eospalax fontanieri*) and myospalacinae (*Nannospalax ehrenbergi galili*) diverged during the early myocene. Their common ancestor likely lived in a similar warm and humid environment.

T4: The transition is estimated to have occurred in the *Saccostomus* group. Given the salt and pepper pattern in its group and the *Steatomys* sister group (green arrows on supplemental Data S4), mesic environment was considered likely ancestral for *Dendromus melanotis* (not a reversion).

T5 and T6 are unambiguously independent transitions, with basal mesic species or outgroups (green arrows on supplemental Data S4).

T7 and T8: unambiguous recent independent transitions in a clear mesic context.

T9: the transition is estimated to have occurred at the basis of the *Mastomys* group, with clear mesic sister groups (green arrow on supplemental Data S4).

T10 and T11: the ancestral state reconstruction of the *Mus* group is perturbed by badly resolved nodes in the Fabre phylogeny (black arrow, nodes in the red ellipse on supplemental Data S4). The evolution of this group is however well known and rather involves a mesic ancestor with two independent transitions in the *Nannomys* group (*Mus Nannomys indutus*) and the *Mus* group (*Mus Mus macedonicus)* following recent aridification (Hardouin et al.2024 and Bryja et al 2014).

T12: unambiguous recent transition

T13: the ancestral state reconstruction slightly favored a single transition for *Mesocricetus* and *Cricetulus* groups. This was surprising since there are mostly arid species in Cricetinae, suggesting that the trifurcation at the basis of Cricetinae may have perturbed the reconstruction. We thus considered this single transition was likely correct.

T14: the automatic ancestral state reconstruction estimated that *Peromyscus maniculatus* is a reversion, but phylogeographic analysis of the diversification of the *Peromyscus* group in North America (Castaneda-Rico et al 2025) argues for a mesic ancestral state, with many independent transitions in this group. Because aridification drove speciation in this group, arid species now dominate over mesic species in the tree, biasing the ancestral reconstruction.

T15 and T16: The automatic ancestral reconstruction estimated that the Heteromyidae common ancestor was arid, with a single transition for *Dipodomys ordii* and *Chaetodipus Baileyi*, and a reversion for *Heteromys desmarestianus*, but this was in disagreement with the literature. According to (Hafner et al. 2007), the semi-tropical heteromyinae group (eg. *Heteromys* in our dataset) remained confined to mesic environments, keeping an ancestral bauplan and poorly diversifying, while 4 other groups, out of 2 are of interest in our case: Dipodomyinae (inc *Dipodomys* in our dataset) and Perognatinae (inc *Chaeotodipus* in our dataset), were extensively diversifying, while they were invading northern zones of America, where the major climate change and tectonic events of mid-miocene had caused a shift to cold-arid environment. Further adaptation to arid environments (not only cold ones) occurred with more recent climatic and geologic changes during the pliocene and pleiostocene. As a consequence of this phylogeographic history, the Heteromyidae group is severely unbalanced with a single group of mesic species and 4 groups of arid species. Therefore the state of the ancestral node was corrected manually, and *Dipodomys* and *Chaetodipus* are considered convergent adaptations to arid environments, with *Heteromys* as a control mesic species, as in other studies (Marra, Romero, and DeWoody 2014).

T17: Unambigous transition

T18: Unambigous transition

T19 and T20 : This transition is ambiguous. Either *Heterocephalus*/*Fukomys* may represent the same adaptation to the arid environment in the common ancestor to all Bathyegidae (36 to 29 MYA), where all members except *Heliophobus* have adapted to arid environments (T19/20). Heliophobius would then have reverted to mesic environments. Or, *Heterocephalus*/*Fukomys* may represent two independent adaptations, one in the genus Heterocephalus (T19) and in the lineage of *Fukomys-Bathyergus* (T20), after the divergence of *Heliophobius* 13.4 MYA (see a and b in the Supplemental Table S5). We used **two independent transitions** for phylogenetical analyses. Note that this choice does not impact the results of our species pairs analysis, since those are ancient transitions in all cases.

T21 and T22. Ancestral state reconstruction would slightly favor a scenario with a reversion for *Spermophilus tridecemlineatus*. However, literature argues for two independent transitions, in *Spermophilus/urocitellus parryii* and in *Spermophilus dauricus* from paleotropical ancestors (Zelditch et al. 2015). Here again, reconstruction is likely biased toward estimating arid ancestors because mesic (tropical forest) species have slowly diversified compared to arid (grassland) species. These species were only used for sequence analysis.

Age of transitions.

Only the 17 transitions in the expression dataset were dated. Minimum and maximum ages of transitions to the arid environment in the expression dataset were estimated from the literature and the above annotated transitions (Fig S1 and supplemental Data S4). A bibliography of 18 articles is cited in Supplemental Table S5). Transition dates are expected to be older than the node corresponding to the most basal divergence within a family of arid species but younger than the node linking this family with a family of mesic species. The brackets are represented in the chronogram, Supplemental Fig. S2.

Supplemental method: dissection, sequencing

Kidney dissection

To homogenize dissections between the different collectors, we set up a specific protocol. The main objective was to avoid introducing any bias in gene expression by recovering RNA from subparts of the kidney that would not be representative of the whole organ, or by co-preparing other tissues, such as adrenal gland or fat, with the kidney. Animals were mostly captured during the night or early morning and killed using cervical dislocation for small animals and a lethal intracardiac dose of pentobarbital for bigger animals administered under deep anesthesia. Immediately after, the kidneys were dissected. Adrenal glands were carefully removed as well as fat using a stereomicroscope when available. Dissections were carried out in a petri dish placed on ice, with cold cell culture medium, or PBS or HBSS solution. Kidneys were then transferred in a small cell culture dish with RNA later (THERMOFISHER – AMBION solution, AM7020) and cut in small pieces of approximately 2-3mm^3. The pieces with the RNA later were then transferred to 2 mL (or 14 mL depending on the size of the kidney) tubes with at least 5-10 volumes of RNA later. When possible, tubes were agitated overnight at 4°C on a rocker and then stored at -20°C. For field captures, samples were occasionally kept at 4°C for 1-2 days.

RNA extraction and sequencing

We prepared RNA-seq libraries for 57 samples (Supplemental Tables S1, S2). For representativity, we used the whole kidney, including for large-sized species. All pieces from a single kidney were lysed in TRIzol with a Precellys homogenizer (Bertin). When needed, several lysates were prepared independently, and then carefully mixed together to ensure homogeneity of the lysate before precipitation and further purification using the RNeasy mini kit from QIAGEN. RNA quantities varied between 250ng/ul and 1800ng/ul, consistent with large kidney size differences. RNA integrity was controlled on a Tapestation (Agilent Technologies), most samples had a RIN between 7.8-10, 5 samples had a RIN between 6.5 and 7.1 but were nevertheless selected. poly(A)+ libraries of the large-scale dataset were prepared with the TruSeq V2 kit (Illumina, non-stranded protocol), starting with 150 ng total RNA and performing only 12 cycles to minimize amplification biases. Libraries were sequenced (Illumina HiSeq 4000, 100bp paired-end or 50bp single-end reads, see Supplemental Table S2). We evenly distributed 10 samples on 5 lanes for single-end sequencing and 6 samples on 4 lanes for paired-end sequencing (resequencing of one library per species). Only single-end sequencing data were used for expression analyses, while paired-end data were used to assist transcriptome assembly.

Supplemental Data

Supplemental Data S1: Phylogeny of our species with the Total_seq sequences.

Phylogenetic tree used for Fig. 2 and for detecting convergent sequence evolution. It was generated from amino acid alignments of the 4,065 complete gene families (as described in the Method section “Phylogenetic reconstruction”). The tree is provided in a newick (nw) format.

Supplemental Data S2: Complete phylogeny with bioclimatic variables.

Bioclimatic information (BIO12 et BIO14) were added onto the phylogeny published by Fabre et al. (2012). Arid (orange), mesic (green) and super-mesic (blue, only for Bio12) categories are indicated.

Supplemental Data S3: Phylogeny of our species with bioclimatic variables.

We subsetted the nodes corresponding to our dataset from the published phylogeny (Fabre et al. 2012) and added bioclimatic information (BIO12, BIO14 and BIO17). We used the median values for each species without filtering (left), filtering out samples from zoos, museums or laboratories (middle), and using the geographic location of the newly sequenced species (in this study)(right). Values per month are shown in histograms.

Supplemental Data S4: Complete phylogeny with estimated ancestral states.

Ancestral reconstructions were carried out by using maximum likelihood on 1898 species from the rodent tree by (Fabre et al. 2012). Colors represent arid (orange) and mesic (green) states at the tip of the branches, and pie proportions represent scaled likelihoods of each ancestral state.

Supplemental Figures and tables.

Supplemental Table S1: List of species used in the study (related to Fig. 2).

General information including BIO17 values, biological status (arid or mesic), Ensembl and NCBI reference for cDNA and RNA-seq data respectively, and whether the species is used in the different datasets of the study.

Supplemental Table S2: List of newly sequenced RNA-seq samples produced in this study.

Information related to the samples collected and sequenced in this study, and submitted to the European Nucleotide Archive (ENA) under accession number PRJEB54931. ‘SAMPLE_NAME’ corresponds to the name we used from RNA extraction to data analysis. ‘SAMPLE_NAME_PROVIDER’ is the name of the sample from the providers. For each species, one sample was sequenced twice (PE and SE) to optimize *de novo* assemblies. Information related to capture include the provider, whether samples were lab grown or field captured and the city of capture. We indicated their use either to build the assembly and/or to perform expression analyses. Note that some samples were not used in our study but we submitted the data.

Supplemental Table S3: Complete list of RNA-seq samples and related RNA-seq information.

This list contains SRA samples downloaded from NCBI (rodent Illumina RNA-seq query, indicated as external in the column ‘SOURCE’) and newly sequenced RNA-seq samples (indicated as internal in ‘SOURCE’). ‘ID_FINAL’ corresponds to the short name used in the study. We associated a ‘BATCH_NUMBER’ to each ‘BIOPROJECT’. Sequencing information include Illumina version, and sequencing strategy (Single or paired end, and average read length). Post-sequencing quality information include GC content, number of total reads (spots for PE) and percent of reads identified (mapping against a known genome). We indicated their use either to build the assembly and/or to perform expression analyses. For each dataset, "group_batch" columns correspond to the groups (indicated as upper letter) used to apply ComBat-Seq batch correction as well as the DE permutation statistics.

Supplemental Table S4: Complete table with Bioclimatic variables.

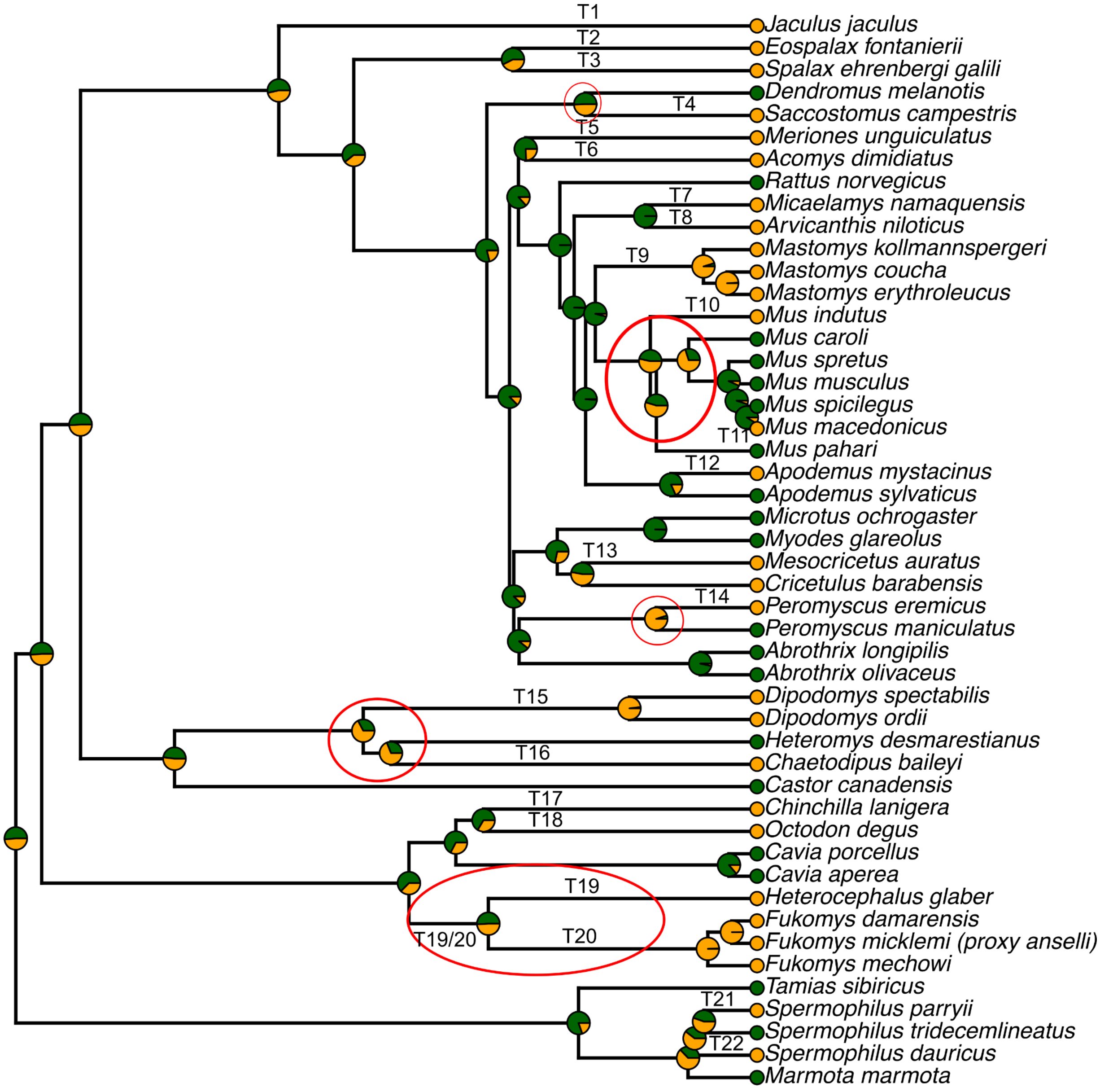
Bioclimatic values were downloaded from the international database worldclim including median temperature per month, median precipitation per month, sum precipitation per month and all 19 categories (BIO1 to BIO19 described here: https://www.worldclim.org/data/bioclim.html). They were extracted using dismo package (see main Method section “Bioclimatic variables”) for all rodent species of the phylogeny published by Fabre et al. (2012). Values were calculated with and without samples from zoos, museums or laboratories.

Supplemental Table S5: Minimum/maximum ages of transitions to the arid environment.

(a) and (b) represent 2 possibilities, see supplemental Methods “Age of transitions”.

Taxon	Interval (MYR)	Justification	Reference
Peromyscus eremicus T14	4.4-3.2	Between the divergence of the species groups eremicus/maniculatus and the diversification of the *eremicus* species group.	(Castañeda-Rico et al. 2025)
Mesocricetus auratus T13	19.6-15.2	Between the divergence of Cricetinae/Arvicolinae and the diversification of Cricetinae	(He et al., 2020)
Mus macedonicus T11	1.15-0.6	Between the divergences *Mus macedonicus/M. spicilegus* and *M. macedonicus/M. cypriacus*	(Hardouin et al. 2024)
Mus indutus T10	3-1	Starting with the divergence *Mus indutus/Mus minutoides* during a period of environnemental changes and expansion of open savannah habitats	(Bryja et al. 2014)
Mastomys 19	5.5-4	Between the divergence Mastomys/Stenocephalemys and the diversification of Mastomys	(Tatiana Aghová et al. 2018)
Apodemus mystacinus T12	7.2-4.5	Beween the divergences of *A. sylvaticus/A. mystacinus* and of *A. mystacinus/A. epimelas*	(Darvish et al., 2015)
Micaelamys namaquensis T7	7-5.8	Between the divergences *Micaelamys/Aethomys* and of *M. namaquensis/M. granti*	(Mikula et al. 2021)(Bothma et al., 2019)
Arvicanthis niloticus T8	<0.5	Divergence *A. niloticus/A. abyssinicus*	(Mikula et al. 2021)
Acomys dimidiatus T6	12-8.7	Between the divergence Acomys/Lophuromys and the diversification of Acomys	(Tatiana Aghová et al. 2018) (T. Aghová et al. 2019)
Meriones unguiculatus T5	18.1-10.8	Between the divergence Gerbillinae/Deomyinae and the diversification of the Gerbilllinae	(Tatiana Aghová et al. 2018)
Saccostomus campestris T4	9-3.9	Between the divergence Saccostomus/Cricetomys and the diversification of Saccostomus	(Mikula et al. 2016)
Eospalax fontanierii T2	25.1-22.4	Between the divergence Myopalacinae/Rhizomyinae and the divergence of Myospalacinae/Spalacinae	(He et al., 2020)
Chaetodipus baileyi T16	20.5-16.2	Between the divergence Perognathinae/Heteromyinae and the diversification of Perognathinae	(Upham, Esselstyn, and Jetz 2019);

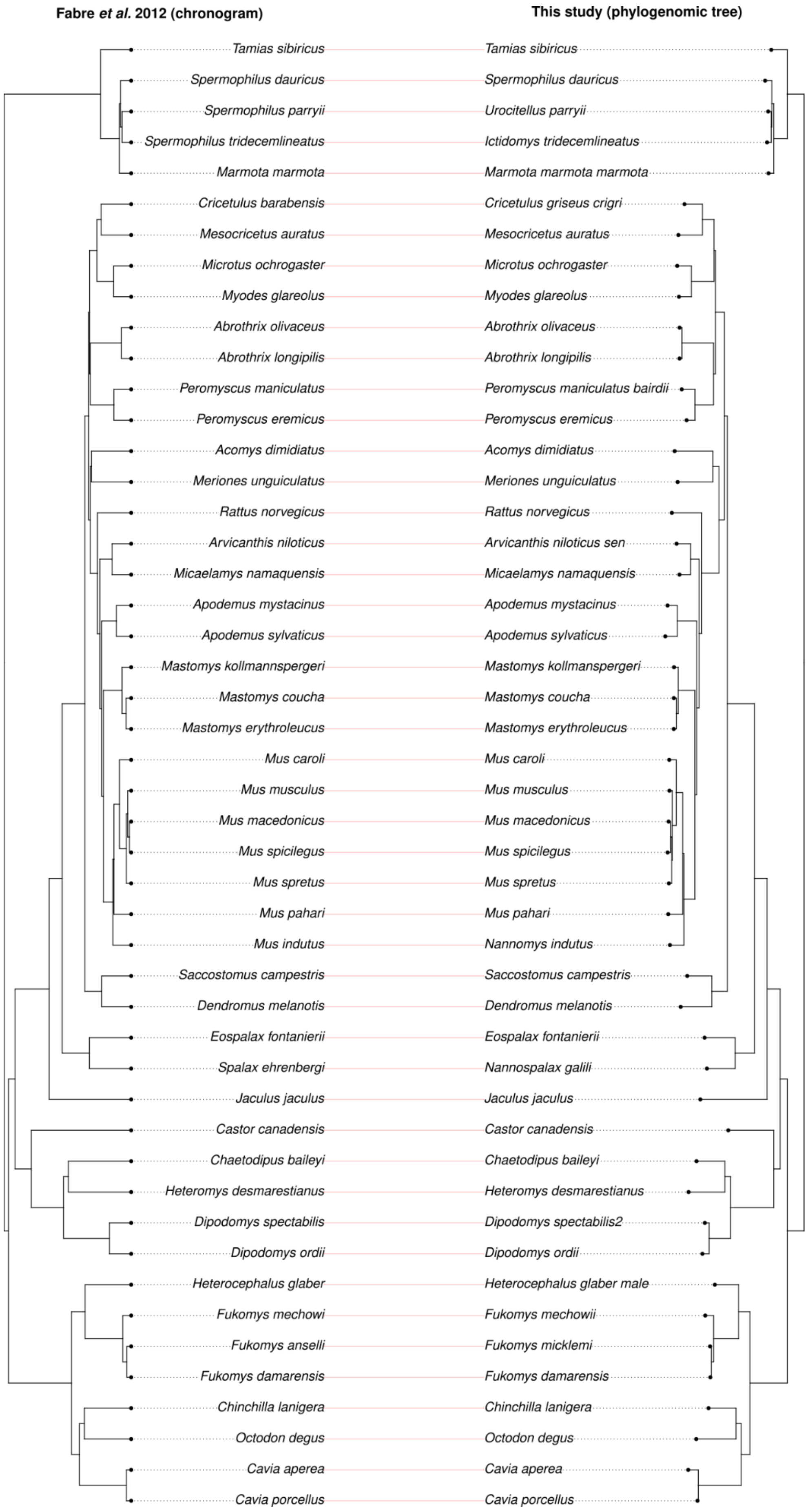
Dipodomys spectabilis T15	25.6-13.5	Between the divergences Dipodomyinae/Heteromyinae and the diversification of Dipodomyinae	(Upham, Esselstyn, and Jetz 2019; Hafner et al. 2007);
Heterocephalus glaber (a) T19	after 29	After the divergence of Heterocephalus/other Bathergidae. Heterocephalus is a monospecific genus, therefore it is difficult to find a minimum age.	(Uhrova et al., 2022)
Fukomys (a) T20	13.4-10.6	Between the divergence Fukomys/Heliophobius and the divergence of Fukomys and Bathyergus	(Uhrova et al., 2022)
Batheyrgidae Fukomys+ Heterocephalus (b) T19/20	36-29	Between the divergence of Bathyergidae/Thryonomyidae and the divergence Heterocephalus/other Bathergidae	(Swanson, Oliveros, and Esselstyn 2019)(Uhrova et al., 2022)
Chinchilla lanigera T17	19.7-7.4	Between the divergences of Chinchilla/Dinomys and Chinchilla/Lagidium	(Upham and Patterson, 2015)
Octodon T18	Not determined		(Upham and Patterson, 2015)
Urocitellus and Spermophilus T21 and T22	Not determined		(Zelditch et al. 2015)
Jaculus T1	Not determined		(Pisano et al., 2015)



Supplemental Figure S1: Ancestral reconstruction of arid and mesic environments.

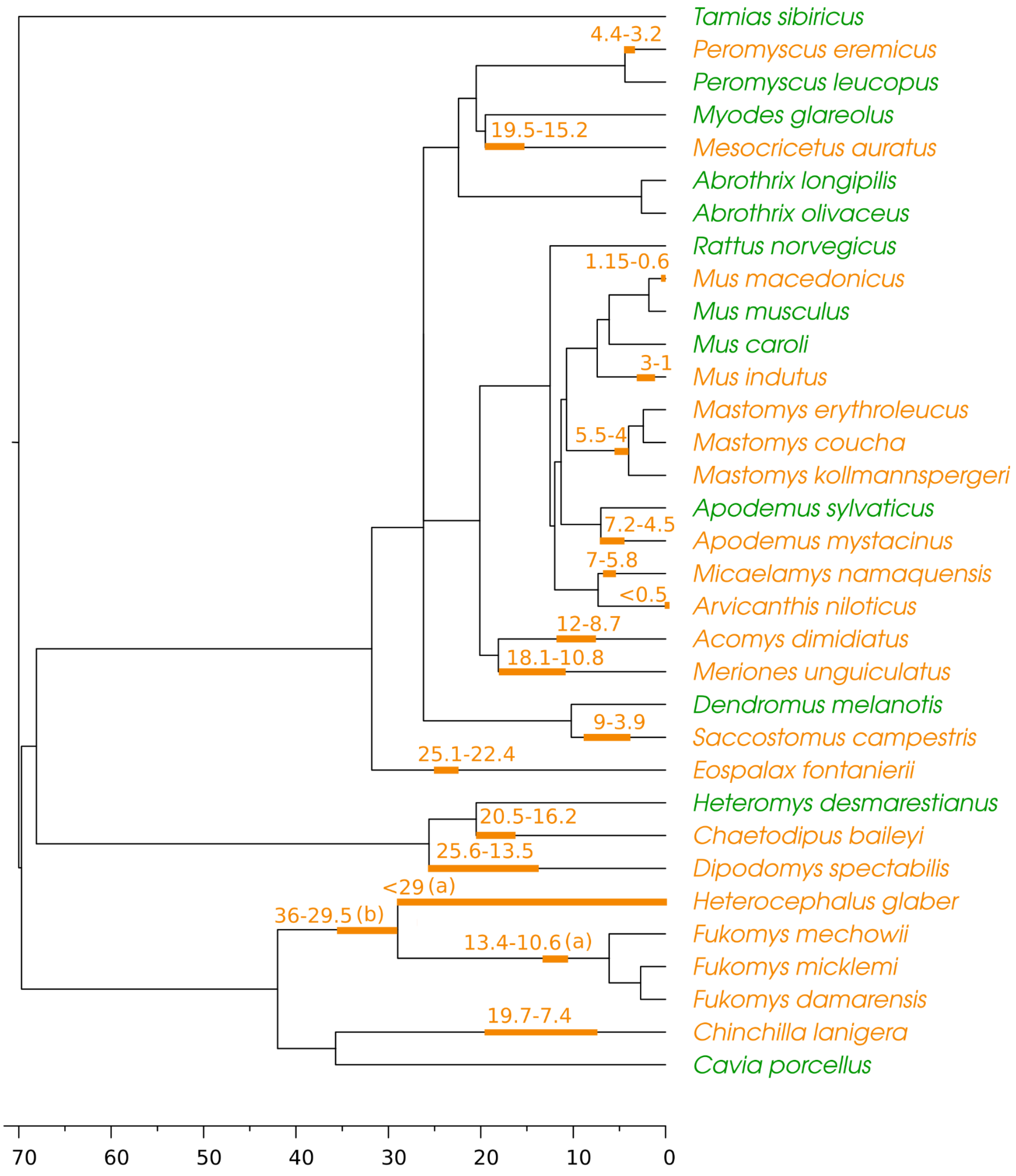
Ancestral reconstructions were carried out by using maximum likelihood on 1898 species from the rodent tree by (Fabre et al. 2012). We then extracted the ancestral states for the subset of nodes corresponding to our dataset. Colors represent arid (orange) and mesic (green) states at the tip of the branches, and pie proportions represent scaled likelihoods of each ancestral state.

Transitions were manually curated (see supplemental Data S4) based on this automatic reconstruction and the literature, resulting in the 22 transitions of the dataset Total_seq. Discrepancies between the manual reconstruction and the literature are shown with red circles. All transitions are discussed in the supplementary text. All Related to Fig. 2


Fabre et al. 2012 (chronogram)
This study (phylogenomic tree)
Tamias sibiricus
Spermophilus dauricus
Spermophilus parryii
Spermophilus tridecemlineatus
Marmota marmota
Cricetulus barabensis
Mesocricetus auratus
Microtus ochrogaster
Myodes glareolus
Abrothrix olivaceus
Abrothrix longipilis
Peromyscus maniculatus
Peromyscus eremicus
Acomys dimidiatus
Meriones unguiculatus
Rattus norvegicus
Arvicanthis niloticus
Micaelamys namaquensis
Apodemus mystacinus
Apodemus sylvaticus
Mastomys kollmannspergeri
Mastomys coucha
Mastomys erythroleucus
Mus caroli
Mus musculus
Mus macedonicus
Mus spicilegus
Mus spretus
Mus pahari
Mus indutus
Saccostomus campestris
Dendromus melanotis
Eospalax fontanierii
Spalax ehrenbergi
Jaculus jaculus
Castor canadensis
Chaetodipus baileyi
Heteromys desmarestianus
Dipodomys spectabilis
Dipodomys ordii
Heterocephalus glaber
Fukomys mechowi
Fukomys anselli
Fukomys damarensis
Chinchilla lanigera
Octodon degus
Cavia aperea
Cavia porcellus
Tamias sibiricus
Spermophilus dauricus
Urocitellus parryii
Ictidomys tridecemlineatus
Marmota marmota marmota
Cricetulus griseus crigri
Mesocricetus auratus
Microtus ochrogaster
Myodes glareolus
Abrothrix olivaceus
Abrothrix longipilis
Peromyscus maniculatus bairdii
Peromyscus eremicus
Acomys dimidiatus
Meriones unguiculatus
Rattus norvegicus
Arvicanthis niloticus sen
Micaelamys namaquensis
Apodemus mystacinus
Apodemus sylvaticus
Mastomys kollmanspergeri
Mastomys coucha
Mastomys erythroleucus
Mus caroli
Mus musculus
Mus macedonicus
Mus spicilegus
Mus spretus
Mus pahari
Nannomys indutus
Saccostomus campestris
Dendromus melanotis
Eospalax fontanierii
Nannospalax galili
Jaculus jaculus
Castor canadensis
Chaetodipus baileyi
Heteromys desmarestianus
Dipodomys spectabilis2
Dipodomys ordii
Heterocephalus glaber male
Fukomys mechowii
Fukomys micklemi
Fukomys damarensis
Chinchilla lanigera
Octodon degus
Cavia aperea
Cavia porcellus

Supplemental Figure S2: Cophylogeny between a chronogram from the literature and our phylogenetic tree.

Left: Tree extracted from a large chronogram with 2260 rodent species (Fabre et al. 2012) that was inferred based on a smaller number of genomic regions (6 mitochondrial genes :12S rRNA, D-Loop, COX3, Cytb, NADH4, NADH1 and 5 nuclear genes : BRCA1, IRBP, GHR, RAG1, vWF). Right: Tree made in this study using phylogenomic data. The topologies are perfectly congruent except the relative branching order of *Mus spretus* and *Mus musculus.* This is not surprising as our *Mus musculus* samples are lab mice, known to have hybrid origins. Of note, *Cricetulus griseus* and *Cricetulus barabensis* are considered subspecies by some in the literature and we choose to join them in the trees. Related to Fig. 2



Supplemental Figure S3: Age of transitions to an arid environment.

Chronogram for the 33 rodent species included in the Total_expression dataset. Dated molecular phylogeny obtained from TimeTree (Kumar et al. 2022) and modified with data in Supplemental Table S5. Colors represent arid (orange) and mesic (green) states of the species. Justifications for the dates of transitions to arid environments in Supplemental Table S5. (a) and (b) represent 2 possibilities, see Supplemental Methods "Age of transitions". Related to Fig. 2

Appendix 2. Convergent expression levels in the Total dataset and in the Murinae subfamily

Supplemental Method: Sample selection and QC

Choice of samples and quality checks:

To keep a proper comparison in the analyses, we performed de novo assemblies for all species investigated at the expression level. We used the following criteria to select samples:

- Priority to Paired-End sequenced samples and reads of longer size
- Homogeneity of the number of reads/spots between species
- Good BUSCO scores

All these features are available in Supplemental Tables S3 and S13.

We selected a maximum of three individuals per species to build our expression-related datasets using the following criteria:

- Homogeneity of the samples belonging to a species, i.e. all samples of a given species must be SE or PE
- Robust species identification, particularly regarding the newly sequenced samples. For example, we removed *Mus Nannomys* collected in Cotonou (ERR100344863, ERR10034836) for uncertain species assignment.
- Homogeneous GC% within a species, and best %reads mapped to a known reference genome were used as a third discriminating criteria.

All these features are available in the Supplemental Table S3.

For nine species, we only secured one replicate. But in most cases, closely related species from the same genus can serve as biological pseudo-replicates for the considered environmental transition.

Quality checks :

To ensure the absence of systematic bias between arid and mesic conditions, we first assessed the technical quality of sequencing data. Quality metrics including average read length, total number of sequenced reads, percent of reads identified (mapping against a known genome) and GC content show no difference between arid and mesic samples (Supplemental Fig. S5A-B). PCA using corrected count data revealed that samples clustered by biological condition (arid/mesic) on the fifth component, with no apparent batch effects or condition-specific biases when colored by sample origin (bioproject batch), average read length, or RIN assessment (Supplemental Fig. S5C-F).

Choice of species in subsets:

For the Murinae subset, we selected Murinae species from our dataset. *Arvicanthus niloticus and Micaelamys namaquensis* represent two independent transitions, but have no neighboring mesic species. and our dataset was desequilibrated in favor of arid species, we decided to remove *Arvicanthus niloticus* (1 sample) for equilibrating the number of mesic (6) and arid (8) species in the dataset. This resulted in 30 samples for the expression dataset, with 5 transitions to the arid status.

For the “recent” and “ancient” subsets, we dated the transitions to arid condition. We built the two datasets with an equal number of samples, 31, and an equal number of transitions to the arid status, 4.

Supplemental Method: phylogenetic PCA

Phylogenetic PCA (Jombart et al. 2010; Jombart et al. 2010) is a multivariate analysis specially dedicated for data with phylogenetic structure. This method uncovers the main structures observable in multivariate data associated with a phylogeny : It identifies principal components in observed trait values in relation with phylogenetic correlation, which is the non-independence of taxa given their phylogenetic proximity. When some of these components display a positive phylogenetic correlation, it means that the trait values in a set of taxa tend to be more similar within closely related taxa. On the other hand, when other components show a negative phylogenetic autocorrelation, it means that distant taxa tend to be closer with respect to a given trait than randomly chosen taxa. It may be the hallmark of convergent evolution.

Supplemental Method and results: EVEmodel and simulations

EVEmodel (Rohlfs and Nielsen 2015) is designed to explore whether a set of branches (foreground) would be up or down regulated in comparison to all other branches (background). They contrast an homogeneous OU model, which uses one limit expression level ’theta1’, with two OU models, which uses two limit expression levels, one for each group (background ‘theta1’ and foreground ‘theta2’).

We downloaded EVEmodel latest implementation from (Gillard et al. 2021) (https://rohlfslab.weebly.com/software.html) and adapted the code so that it could run with a single sample in some species, which is the case in our dataset (the patch is included in our gitlab repository). Then we applied it to look for differences in expression between arid and mesic species. We used norm-transformed counts of expression (method ntd from DESeq2 package), to favor EVEmodel that models normal distribution at leaves (because of OU modeling). For each gene, we performed the ‘twoThetaTest’ method contrasting one homogeneous tendency on all branches, or two phenotype specific tendencies (mesic vs arid).

First, we observed that very few genes passed the FDR threshold (adjusted p-value <0.1). However, among the genes found by DESeq2 (adjusted p-value <0.1), many genes were significant with EVEmodel (p-value < 0.05). The results are integrated into Fig. 3D,H and Fig. 6D,H. To compare the results of both methods, we plot in Supplemental Fig. S12A, the p-value of twoThetaTest for each gene against its log2FoldChange from DESeq2. Genes selected by DESeq2 are in red. We can see in Murinae_expr dataset that DESeq2 selected genes are mostly detected by EVE, but there are also large differences, where genes with very large log2FoldChange are not detected by EVE. On the contrary, EVE detects many genes with very small log2FoldChange, but since those p-values are not corrected for multiple tests, there may be many false positives.

Using p-value < 0.05 threshold, the main disaccordance between DESeq2 and EVEmodel approaches is on the relative numbers of detected genes between Murinae_expr and Recent_expr datasets, with a ratio from 16 to 2. We can expect results to differ between both methods, since they focus on two different features : EVEmodel models the evolution towards two optima against one optimum, while DESeq2, in our case, models expression shifts knowing the family batch effect, even though the shift has several scales among species. Also, EVEmodel significance is related to the difference between theta2 and theta1, where DESeq2 looks for multiplicative changes, which means that its significance is related to the ratio between expression levels. On top of these general features, we performed simulations that show the impact of the tree shape on EVE detection (Supplemental Fig. S12).

Simulations were made on each data set, using the function simTwoTheta, with the true annotated tree, and alpha, beta, sigma parameters set to the median of their respective estimated value on this data set. For each gene we used its estimated theta1, and with a fixed ratio theta2/theta1, or theta1/theta2, depending if the shift was detected up or down. To properly compare the simulations on both datasets, we set the ratio on thetas such that we get similar distributions on observed arid/mesic expressions in both cases (distributions in Supplemental Fig. S12 B,E). Then, we performed the twoThetaTest on both simulated datasets. As we see in Supplemental Fig. S12 C,F), the p-values on Murinae simulated data are much larger that those on Recent simulated data (with a factor 10). This means that on similar expression profiles, detection on Recent (ie. short tree) is much more sensitive than on Murinae.

Also, we can observe that p-value depends negatively on theta1, and that in the Recent dataset down-regulated genes (red points) are much more difficult to detect than up-regulated (blue points). These features make the comparison of results between EVEmodel and DESeq2 even more difficult to interpret.

Supplemental Method: Testing the significance of the number of DE genes found using DESeq2

To estimate whether the number of observed Differentially expressed (DE) genes (found using DEseq2) between arid and mesic species is significantly different from a random observation, we set up a protocol based on permutations inspired by (Bittner, Mack, and Nachman 2022), that respect the phylogenetic groups within which the species labels can be permuted. Indeed, it is important to control that the random assignment of labels preserves the dispersion of the two conditions along the phylogeny and does not group species within close taxonomic families, because this would estimate a number of DE genes between taxonomic families and not between xeric and mesic of species distributed between these families.

For each of the datasets, we defined phylogenetic groups within which the species labels can be permuted (Supplemental Table S6). The following steps were then carried out, and repeated 1000 times for each of the datasets.

Step 1: Within each group, each xeric species was associated with a mesic species randomly (without replacement). Some mesic species may remain unchanged if there are more mesic than xeric species, the reverse is not possible. If, within these pairs of species thus formed, there is a difference in the number of samples per species, the maximum number of samples is chosen randomly (without replacement) so that there is the same number of samples in each species of the pair.

Step 2: The selected samples were extracted from the total count table; an “observed” number of differentially expressed genes was calculated with this reduced table retaining the true labels (adjusted p-value <0.1).

Step 3: To prevent our simulations from being biased by the signal of our actual convergent phenotype, we removed the DE genes identified in step2 and replaced them with genes with an equivalent expression level. To do this, we sorted the genes according to their average expression (basemean) and created 20 quantiles. These genes are therefore duplicated in the count table but they will be permuted independently. This consideration is important because if we do not take this step, we find that the number of DE genes in the simulations and the number of DE genes common to the permuted count table and the reduced count table are positively correlated. Indeed this is expected, if by chance the counts of a DE gene in the count table with the real labels are little permuted, it will remain DE and this is directly linked to the signal induced by our convergent phenotype.

Step 4: For each gene and each previously associated pair of species, we swap the expression values between the xeric and mesic samples with a probability of 0.5. At the end of the process, we therefore swapped on average the expression levels of half of the xeric species with those of a mesic species. Of note, we cannot systematically exchange the species phenotypes because this would amount to doing the same analysis but with the labels reversed. The permuted count table is then used to calculate an “expected” number of DE genes.

At the end of 1000 permutations, we performed a paired Wilcoxon test to compare the distribution of “observed” DE gene counts in the reduced count tables with those “expected” obtained in the corresponding permuted count tables.

Supplemental Methods and results: DE genes in Total_expr and Murinae_expr datasets

Functional enrichments

We computed functional enrichment for Gene Ontology terms using *EnrichGO* function from the ClusterProfiler package (Wu et al. 2021) on lists of differentially expressed genes.

Manually curated list of renal marker genes

We also searched for genes known to be involved in kidney diseases or to be specific of certain cell types. For this, we conducted an extensive literature search to list kidney cell types along renal segments with respect to kidney structure and organization (Bankir and de Rouffignac 1985; Lee, Chou, and Knepper 2015; Chmielewski 2003) and their associated specific markers (Supplemental Table S9).

We included information from literature (Habuka et al. 2014; Strutz and Zeisberg 2006; Brunskill and Potter 2010; Roy, Al-bataineh, and Pastor-Soler 2015; Schlöndorff and Banas 2009;

Stefańska, Péault, and Mullins 2013; Huling et al. 2012; Thiagarajan et al. 2011), single-cell RNA-seq data (Park et al. 2018; Cao et al. 2018) and databases (ESBL database; (Huling et al. 2012); https://proteinatlas.org; https://esbl.nhlbi.nih.gov/Databases/KSBP2/). This search allowed us to fine-tune segment-specific cell types. For example, we retrieved different specific markers of epithelial cells depending on their localization, i.e. glomerulus, PCT, DCT, LOH. The curated marker table includes 217 cell type-associated renal segment specific markers; 179 of these genes were kept since they were available in the Murinae dataset (Supplemental Table S9).

Furthermore, a list of 244 disease genes were retrieved from the OMIM database and (Park et al. 2018), of which 165 were found in the Murinae dataset.

Results of DESeq2 for the Total_expr dataset

Unsurprisingly given its modest size, this group of 41 genes revealed only two overrepresented Gene Ontology (GO BP) terms, “small molecule biosynthetic process” and “animal organ morphogenesis” (adjusted p-values < 0.02). Among these genes, we found two members of the solute carrier (SLC) gene family, *Slc35b4* and *Slc40a1* (Kordonowy and MacManes 2017), which is marginally more than expected by chance (Fisher exact test, p-value = 0.059). *Slc40a1* is an iron exporter previously identified in a dehydration experiment (Kordonowy and MacManes 2017). This set also included 5 genes known as kidney markers or associated with renal diseases: C*asr* which is associated with hypocalcemia and calcium kidney stones (Vezzoli et al. 2011; Hanna et al. 2021), *Ctsh*, *Xpnpep2* (Böttinger 2010), *Fam20a* which is associated with enamel renal syndrome (S. K. Wang et al. 2014) and *Cpne2* which is involved in renal cancer (Zhou et al. 2018).

Results of DESeq2 for the Murinae_expr dataset

We found 17 marker genes and 12 disease genes in the list. Differentially expressed genes included 2 aquaporins (*Aqp2*, a vasopressin-regulated water channel involved in diseases affecting urine-concentrating ability (Pannabecker 2015) and *Aqp7*, expressed in proximal tubules, with phenotypes of insulin resistance and important in glycerol reabsorption in the kidney (Sohara et al. 2006), 17 solute carriers (including the urea transporter *Slc14a2*, the sodium carrier gene *Slc8b1*, *Slc27a2* that plays an important role in hepatic fatty acid uptake and was found overexpressed in kangaroo rat kidney (Marra et al. 2012). Focusing on genes significantly downregulated in xeric species, enriched GO terms included response to insulin, and transmembrane transporters (Supplemental Table S10).

Supplemental Data

Supplemental Data S5: Raw expression count tables.

The raw count table was generated after read mapping using Kallisto, transcript annotation using BLASTX against the EggNOG version 5 database, and using tximport package to scale by the average transcript length. The final count table shows count per gene as explained in the main Method sections.

Supplemental Figures and Tables.

Supplemental Table S6: Number of DE genes with different LFC and padj values.

Comparison of number of DE genes identified using different LFC and different padj thresholds, in the four datasets.

Supplemental Table S7: Differentially expressed genes in all sets, using DESeq2.

The table contains all differentially expressed genes identified in the four datasets, as indicated in column ‘dataset’. Gene MGI are listed in column ‘Gene’ with their associated up- or down-regulation in arid species in column ‘direction’ (corresponding to Abs(L2FC) > 1 & padj < 0.1).

Supplemental Table S8: List of differentially expressed genes in all sets, using EVE model.

EVE likelihood ratio test, parameters (theta1, theta2, sigma2, alpha, beta, direction of shift), and test (log-likelihood of two theta model, likelihood of one theta model, p-value, adjusted p-value) are indicated for each gene.

Supplemental Table S9: List of marker genes.

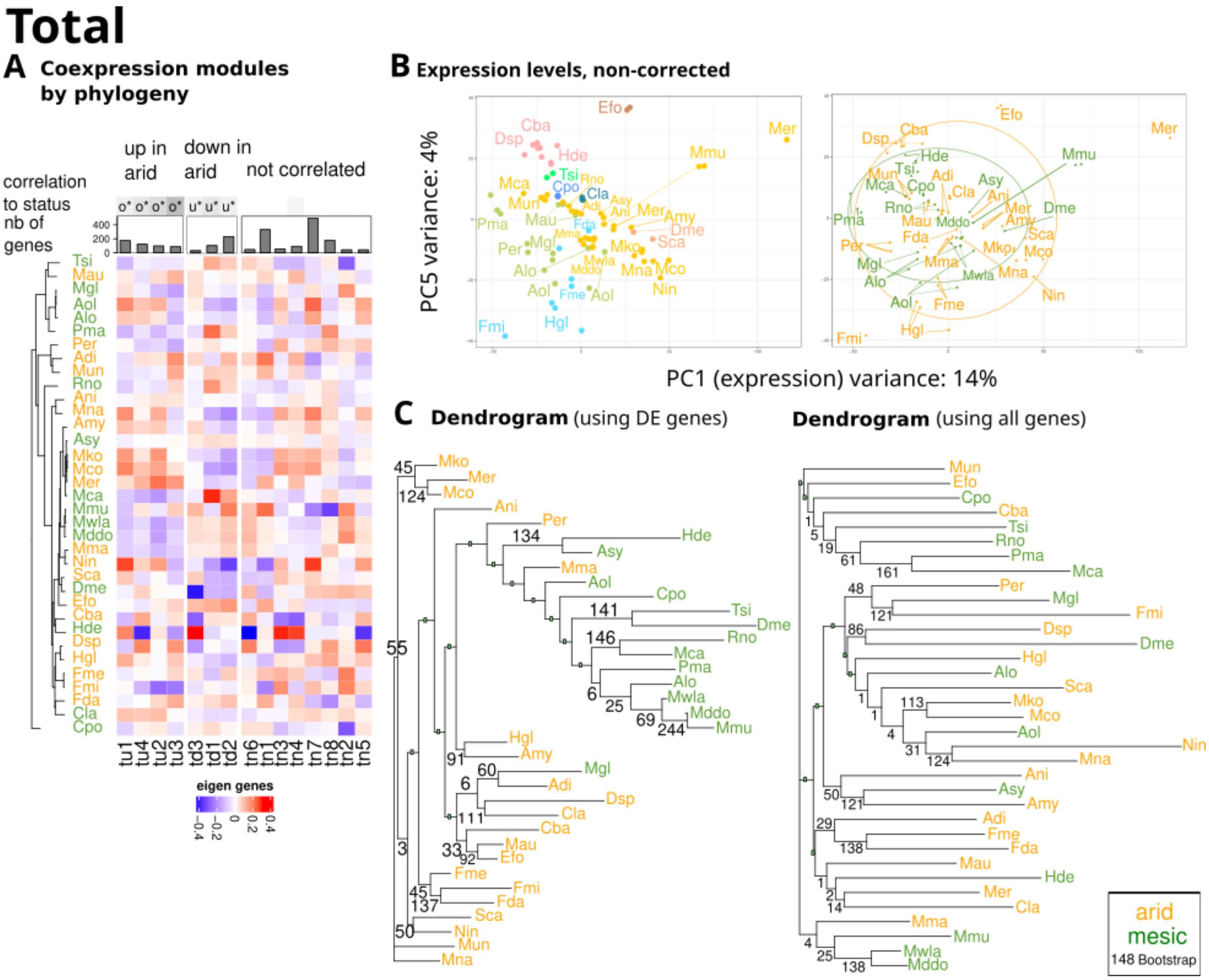
Manually curated list of renal marker genes (as explained in the section ‘Manually curated list of renal marker genes’). Genes can be found in renal segments (e.g. glomerulus, Loop of Henle,...) or specific to cell types (CellType1, 2 and 3 vary in granularity) or associated with renal disease (OMIM). This list was then used to cross-reference findings in DE analyses, and in the tissue composition analyses (Supplemental Fig. S15).

Supplemental Table S10: Functional enrichments for DE and co-expressed modules.

Functional enrichment were performed using ClusterProfiler (EnrichGO function) on lists of differentially expressed genes (Up- and down-regulated together) and co-expressed genes, as indicated in ‘DE_or_WGCNA’, for each dataset (column ‘dataset’).

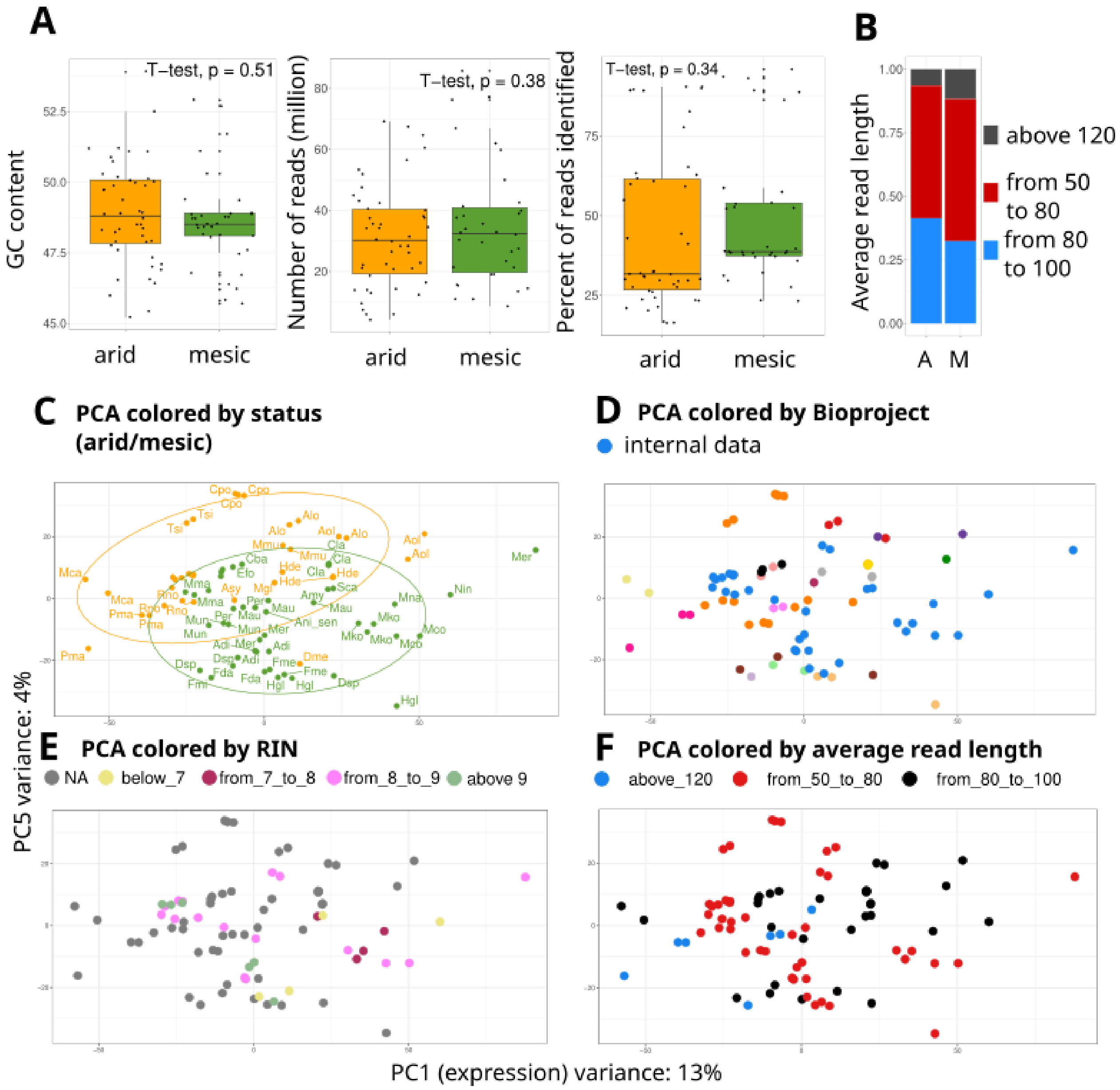
Supplemental Table S11: List of genes for co-expressed modules in all sets.

The table contains two sheets. The first sheet indicates the module name (color-coded), the genes in the module and the dataset. The second sheet indicates the names of modules used in the manuscript associated with its color-coded name, for each dataset.



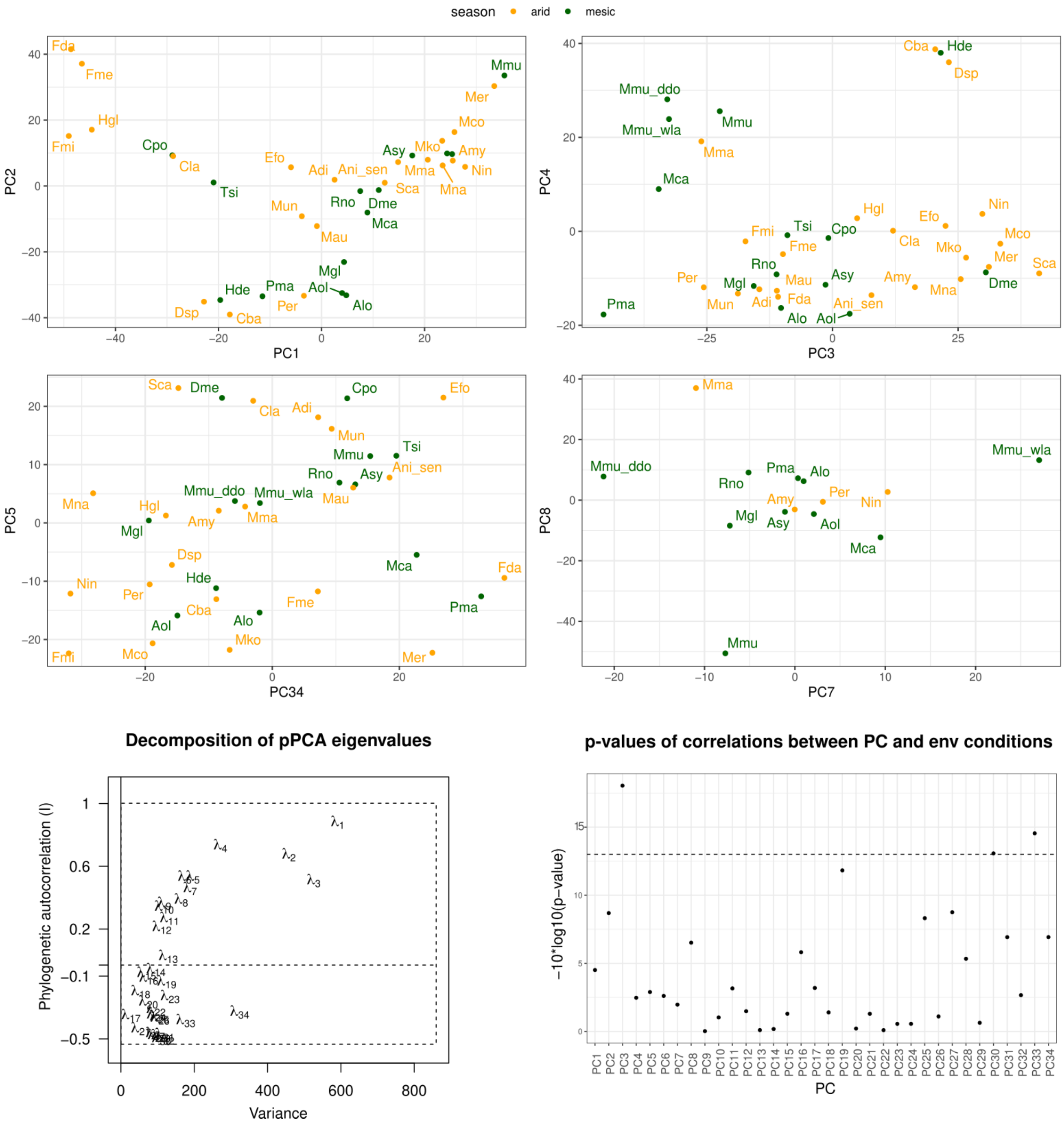
Supplemental Figure S4: Co-expressed modules and dendrograms in Total_expr

A: Co-expression modules represented side by side with the phylogeny show a phylogenetic effect. B. Similarly, the initial components of a PCA of the total dataset separated samples from various rodent families according to the phylogeny, and grouped samples from the same species together, as observed previously in a number of studies (e.g in rodents, Bittner et al. 2022). C: Dendrograms constructed by using neighbor-joining on the basis of expression distance matrices (1 – Spearman’s correlation coefficient) complemented with 1000 bootstrap analyses. All genes (right) and differentially expressed genes (left) were used. Arid and mesic environments in orange and green. Related to Fig. 3.



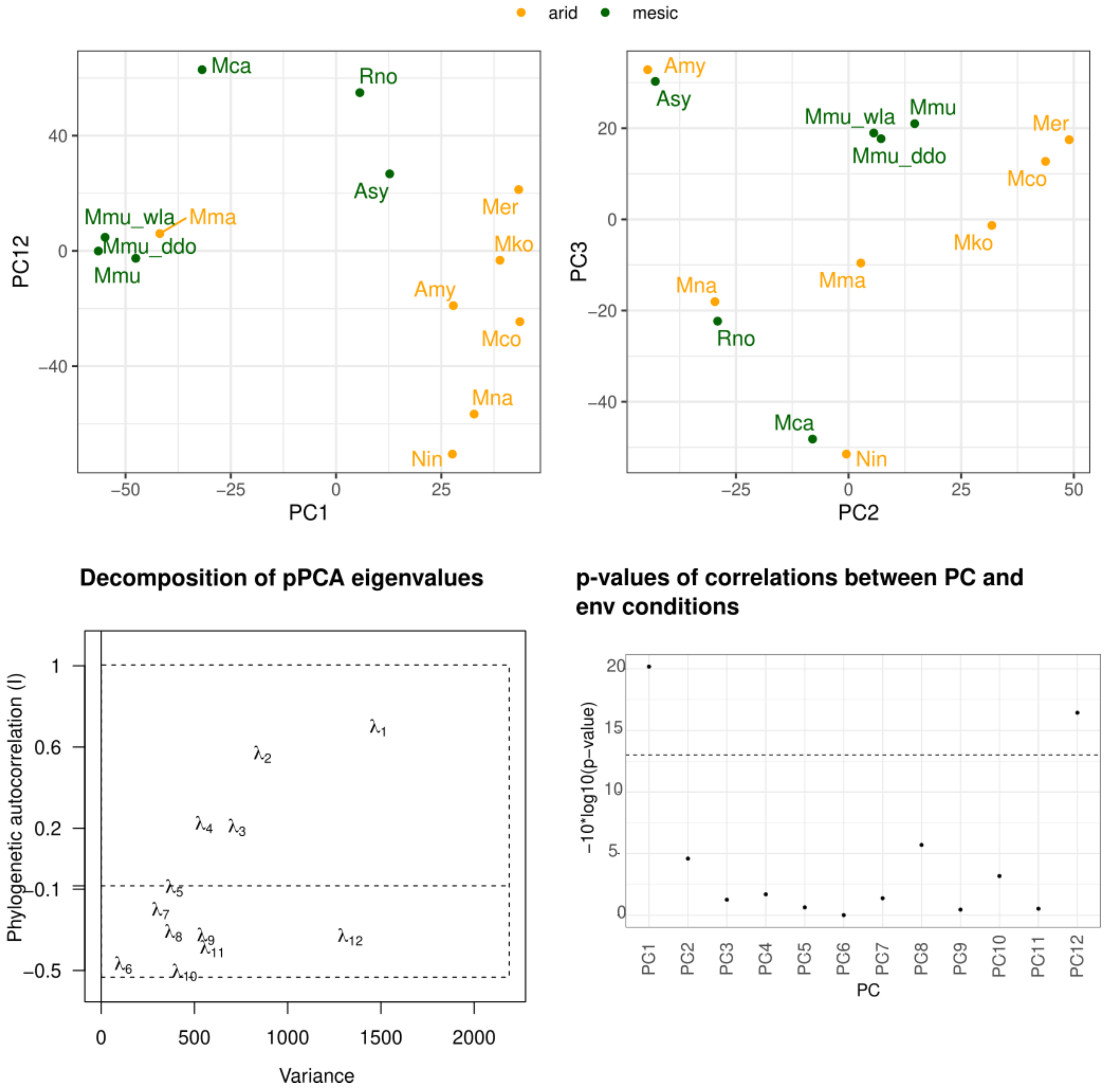
Supplemental Figure S5: Quality control for RNA-seq samples in the Total_expr dataset.

A) Boxplots depicting GC content (left), number of sequenced reads (middle) and percent of reads identified by mapping in known genomes (right). Statistical comparison between arid and mesic were performed using T-test. B) Barplot depicting frequencies average read length between arid and mesic. C-F) PCA plots using family-corrected expression values from the set Total_expr; panel A is colored by environmental status (arid in orange, mesic in green); panel B is colored according to associated NCBI Bioproject with blue being the samples produced in this study (internal) and other colors were retrieved on NCBI; panel C is colored according to RNA quality (RIN before sequencing) only for internal samples; panel D is colored according to average read length using three classes, reads from 50 to 80bp (red), reads from 80 to 100 bp (black) and reads longer than 120bp (blue). Related to Fig. 3.



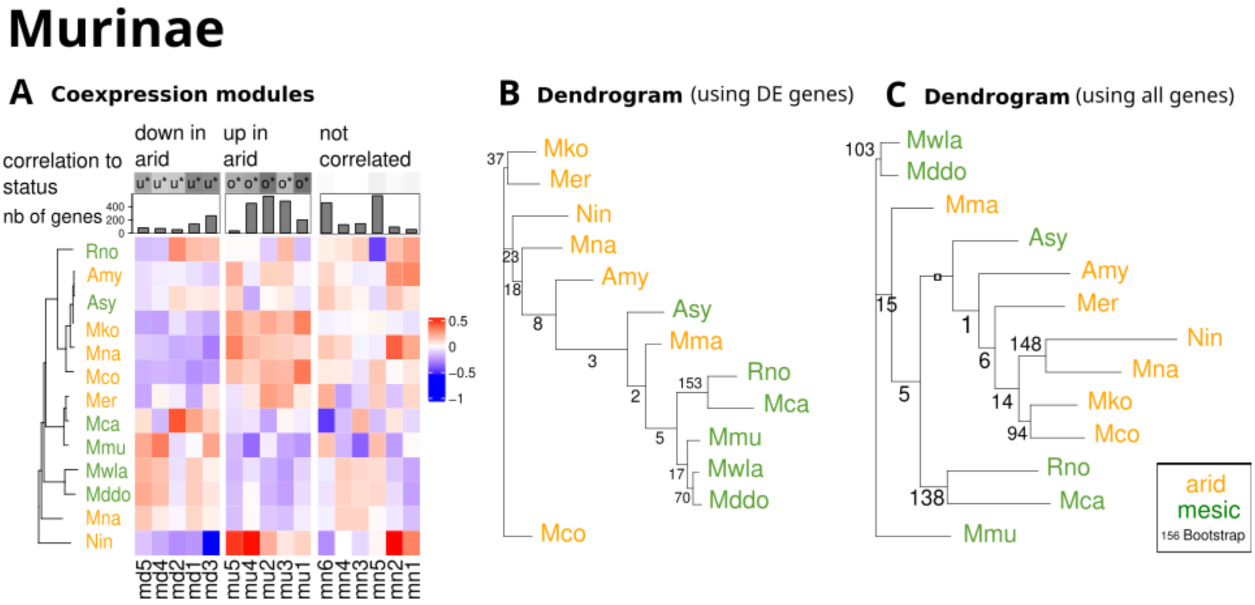
Supplemental Figure S6: phylogenetic PCA (pPCA) for the total dataset.

Top: Species location on the map of the main components of the pPCA. Bottom: Decomposition of the pPCA eigenvalues showing variance, positive phylogenetic autocorrelation (PC1, PC2, PC3, PC4) and negative phylogenetic autocorrelation (PC34). P-values of t.test between PC coordinates and arid/mesic conditions, dashed line represents p=0.05 threshold. Related to Fig. 3.



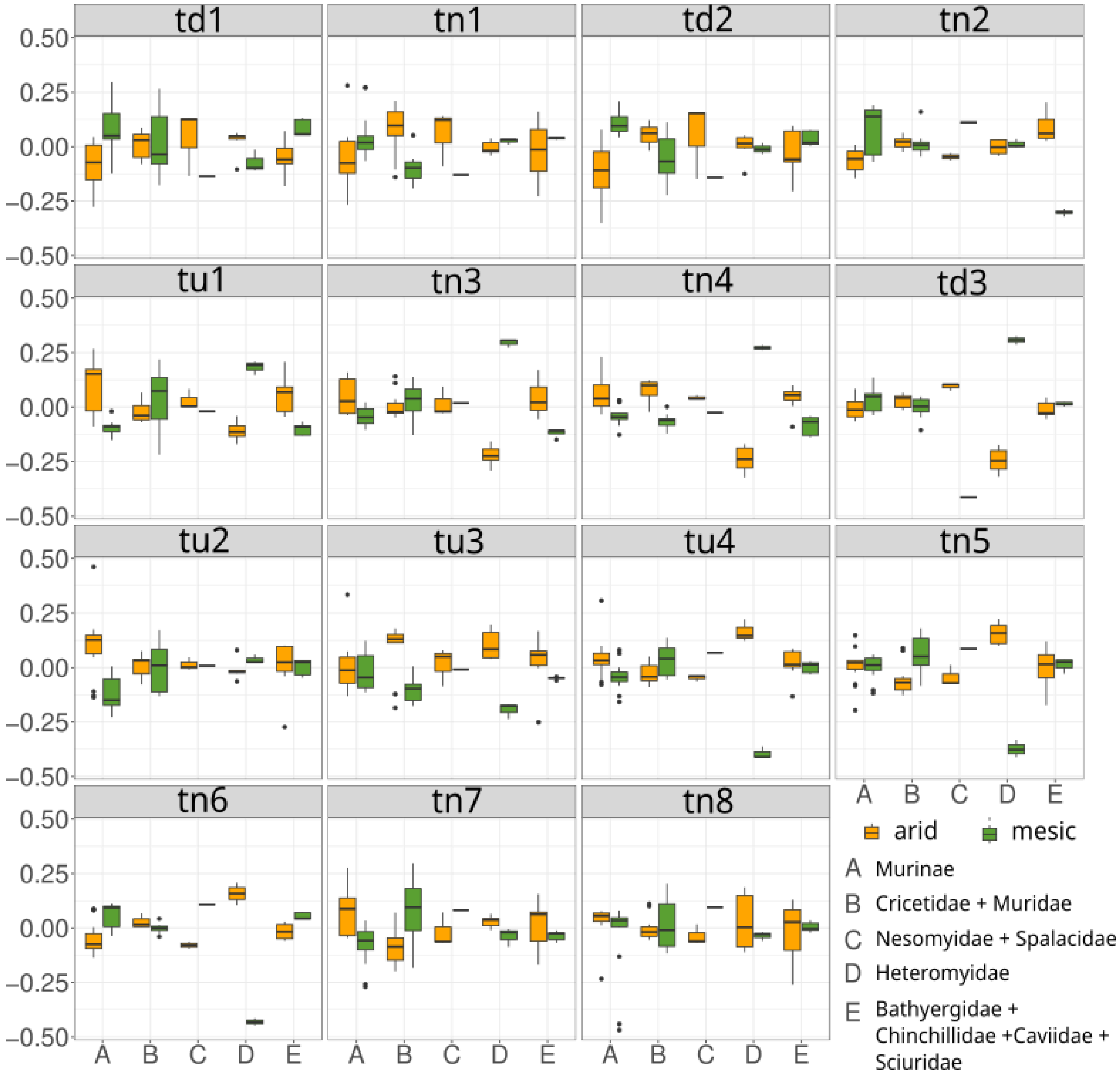
Supplemental Figure S7: phylogenetic PCA for the Murinae dataset

Top: Species location on the map of the main components of the pPCA. Bottom: Decomposition of the pPCA eigenvalues showing positive phylogenetic autocorrelation (PC1, PC2, PC3) and negative phylogenetic autocorrelation (PC12). P-values of t.test between PC coordinates and arid/mesic conditions, dashed line represents p=0.05 threshold. Related to Fig. 3.



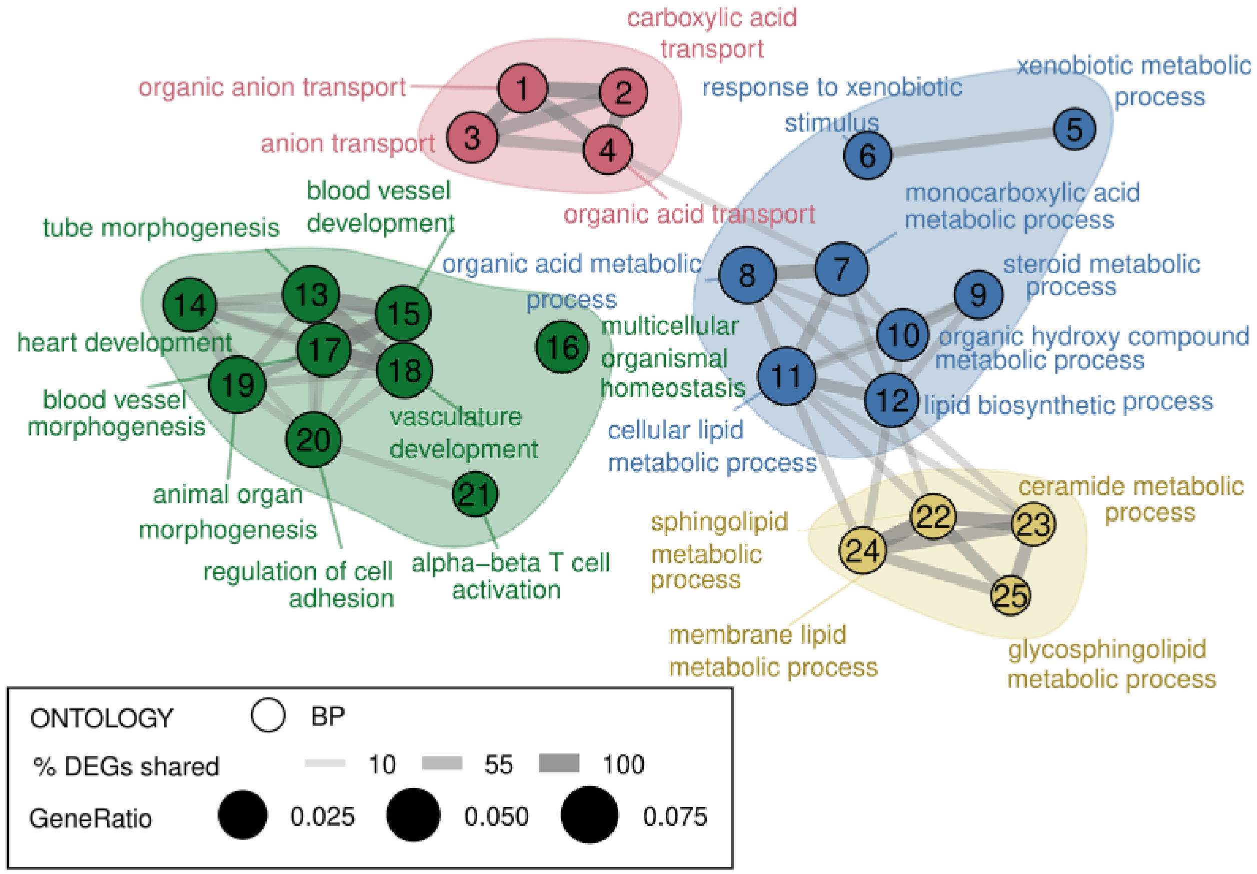
Supplemental Figure S8: Co-expressed modules and dendrograms in Murinae_expr

A: Co-expression modules represented side by side with the phylogeny show a phylogenetic effect ; B, C: Dendrograms constructed by using neighbor-joining on the basis of expression distance matrices (1 – Spearman’s correlation coefficient) complemented with 1000 bootstrap analyses. All genes (C) and differentially expressed genes (B) were used. Related to Fig. 3.



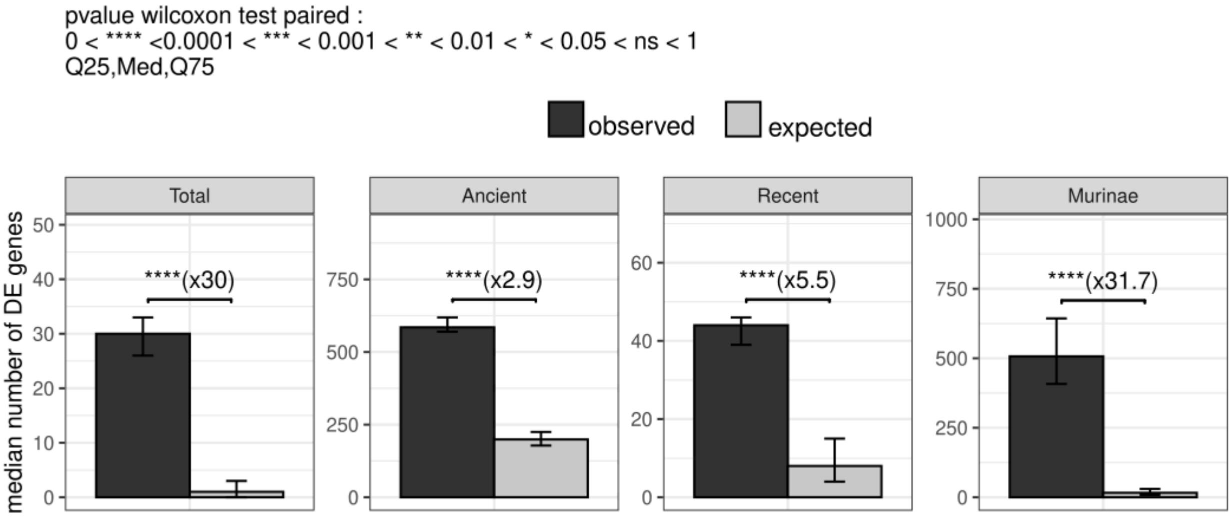
Supplemental Figure S9: Co-expression modules with phylogenetic groups in Total_expr

Co-expression modules from the Total dataset represented as boxplots according to phylogenetic groups. Y-axis values correspond to eigengenes. Module names as in Fig. 3C. (A similar analysis was not made for Murinae_expr because all species belong to the same subfamily). Related to Fig. 3.


carboxylic acid transport
organic anion transport
anion transport
organic acid transport
response to xenobiotic stimulus
xenobiotic metabolic process
monocarboxylic acid metabolic process
organic acid metabolic process
steroid metabolic process
organic hydroxy compound metabolic process
lipid biosynthetic process
cellular lipid metabolic process
blood vessel development
tube morphogenesis
heart development
blood vessel morphogenesis
animal organ morphogenesis
regulation of cell adhesion
alpha–beta T cell activation
vasculature development
multicellular organismal homeostasis
sphingolipid metabolic process
ceramide metabolic process
membrane lipid metabolic process
glycosphingolipid metabolic process
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
ONTOLOGY
○ BP
% DEGs shared
10
55
100
GeneRatio
0.025
0.050
0.075

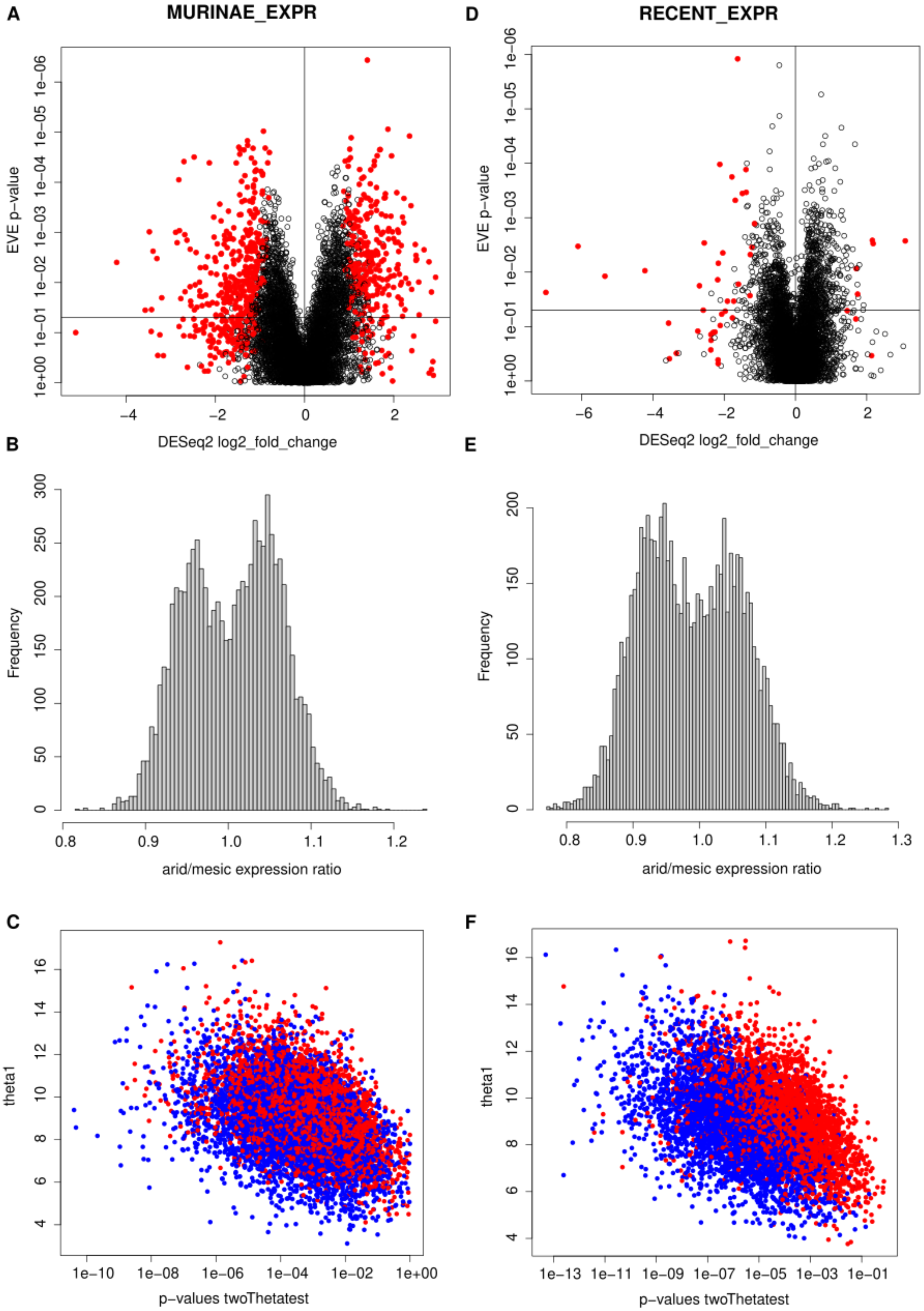
Supplemental Figure S10: Gene Ontology enrichment in modules from Murinae_exp.

Functional enrichments in the 10 modules significantly correlated with the environment in the Murinae_exp dataset (mu and md modules shown in Fig. 3F). Top 25 significantly enriched Biological Process (BP) terms were used (FDR = 0.05), links on the graph represent the percentage of genes in common between two terms. Only edges representing at least 10% of shared genes are drawn. Term names are indicated on the graph. GeneRatio corresponds to genes of interest in the gene set / total genes of interest. Related to Fig. 3.



Supplemental Figure S11: Comparison of true and expected differentially expressed gene numbers found by the DESeq2 method.

The median number of differentially expressed genes (DEG) is shown for scans with true status annotation (observed, black) and for shuffled status annotation (expected, grey). The number of observed DEG is not the same as in the main text because the process involves random sampling and permutations (See supplemental Methods). Confidence intervals and results of Wilcoxon paired tests are indicated. Related to Fig. 3 and 6.


A
MURINAE_EXPR
EVE p−value
1e+00
1e−01
1e−02
1e−03
1e−04
1e−05
1e−06
−4
−2
0
2
DESeq2 log2_fold_change
D
RECENT_EXPR
EVE p−value
1e+00
1e−01
1e−02
1e−03
1e−04
1e−05
1e−06
−6
−4
−2
0
2
DESeq2 log2_fold_change
B
Frequency
0
50
100
150
200
250
300
0.8
0.9
1.0
1.1
1.2
arid/mesic expression ratio
E
Frequency
0
50
100
150
200
0.8
0.9
1.0
1.1
1.2
1.3
arid/mesic expression ratio
C
theta1
4
6
8
10
12
14
16
1e−10
1e−08
1e−06
1e−04
1e−02
1e+00
p−values twoThetatest
F
theta1
4
6
8
10
12
14
16
1e−13
1e−11
1e−09
1e−07
1e−05
1e−03
1e−01
p−values twoThetatest

Supplemental Figure S12: Comparison of genes detected by EVE model and DESEq2 for the Murinae and Recent datasets.

(A,D) The p-value of EVEmodel twoThetaTest for each gene is represented against its log2FoldChange from DESeq2. Genes significant with DESeq2 are in red. The horizontal line indicates the p-value=0.05 threshold. In murinae_expr dataset DESeq2 selected genes are mostly detected by EVE. In recent_expr, EVE detects more genes than DESeq2. (B,E) Simulations by EVE of the expression levels using the parameters estimated previously, with a fixed ratio theta2/theta1, or theta1/theta2, depending whether the shift was detected up or down. To properly compare the simulations on both datasets, the ratio on thetas are set to get similar distributions on observed arid/mesic expression ratio in both cases. (C,F): Result of twoThetaTest on both simulated datasets. The p-values on Murinae simulated data are much larger that those on Recent simulated data (with a factor 10). This means that on similar expression profiles, detection on Recent (i.e. short tree) is much more sensitive than on Murinae. Related to Fig 3 and 6.

Appendix 3. Convergent patterns of tissue composition

Supplemental Method.

Choice of the deconvolution method

Many methodologies to infer proportions of individual cell types from bulk transcriptomics data have been developed, some of which using marker genes for different cell types, and others using scRNA-seq data. We implemented the former using our list of marker genes (see

Appendix 2 and Supplemental Table S9). For the latter methods, we used kidney scRNA-seq dataset from three rodent species for which data from untreated animals is available, mouse, rat and hamster (see below).

To determine the best deconvolution method for our data, we used the available benchmark from Cobos et al. (Avila Cobos et al. 2020). With the best applicability to other Murinae species and good results in *Mus musculus*, MuSiC, a method based on scRNA-seq data, was selected in our analysis (MuSiC_1.0.0, (X. Wang et al. 2019).

Preprocessing of reference scRNA-seq data

Rat data

Rat data was taken from (Balzer et al. 2023) that provides counts for 25,399 genes, 217,132 cells, and 12 individuals. We downloaded this data from GEO (GSE209821) and obtained the raw data matrix post filtering GSE209821_EXPORT_GEO_counts_postfilter.rds. The original study included thorough filtering with the following criteria :

- ambient RNA correction
- doublet removal (see original study)
- removal of nuclei with <200 or >3,000 expressed genes
- removal of nuclei with mitochondrial gene percentages >15

We further selected 48,744 cells corresponding to the 3 lean and untreated individuals. We verified the applied filters (further removing cells with total counts lower than 500 or greater than 10 times the median of counts) using the Seurat package (Seurat_5.1.0(Hao et al. 2021)). The average number of genes expressed per cell was 1325 (median 1271). We verified that the percentage of reads corresponding to mitochondrial genes was moderate. The Supplemental Fig. S13 shows that the 3 individuals differ slightly in terms of number of genes expressed per cell and number of reads per cell (A), but their cell maps are very similar (B). We annotated this UMAP by transferring the annotations from publication (cluster 3 column in the original publication, file GSE209821_EXPORT_GEO_meta.data.csv). We removed immune cells before the deconvolution.

Mouse data

Mouse data was taken from(Park et al. 2018) that provides counts for 16,272 genes with MGI ID, 43,745 cells, and 7 individuals. We downloaded this data from GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE107585) and obtained the raw data matrix GSE107585_Mouse_kidney_single_cell_datamatrix.txt.gz. Quality filtering was already performed by the authors with the following criteria :

- Removal of cells with <200 or > 3000 unique genes expressed (as they are potentially cell doublets)
- Only genes expressed in 10 or more cells are conserved
- Cells with mitochondrial gene percentages over 50% were removed. This filter is quite high compared to other scRNAseq study but the authors justified this in the original paper “the increased mitochondrial gene count was inherent to specific (proximal and distal tubule) cell types in the kidney”.

Data were normalized using SCTransform and UMAP was then generated using 15 dimensions of the PCA (using Seurat_5.1.0(Hao et al. 2021)). We checked these quality filtering by using the standard preprocessing pipeline from Seurat (Supplemental Fig. S14-A). We checked that the cells with high mitochondrial content were not segregated in particular clusters (this is in accordance with the original article “The clustering of cells was not affected by mitochondrial gene content. Fig S2”). Upon reanalysis of these data, we removed one of the 7 individuals. This sample (ind 7) created an additional cluster and lacked several clusters in the published parent study (highlighted Supplemental Fig. S14-B). No additional filtering was performed in our analysis. The average number of genes expressed per cell was 1047 (median 940), the minimum number of reads per cell was 700 (maximum : 15,000). Cell type identities assigned in the original publication (V2 column) were then re-attributed to each cell.

Hamster data

The hamster dataset was taken from an unpublished dataset from *Mesocricetus auratu*s kidney (GEO Accession: GSM4673474, sample SRR12235555). We mapped the data using nf-core/scrnaseq analysis pipeline 2.4.1 (Ewels et al. 2020) to the hamster reference genome Mesocricetus_auratus-GCA_017639785. The dataset contains a single individual and relatively few cells (16378 genes, 2042 cells) with a high mitochondrial content (68% cells with more than 20% mitochondrial reads). We could not therefore use this data for deconvolution, but we nonetheless used it to verify that the marker genes defined in the mouse were expressed in similar cell types in the hamster. For this, we extracted 1,924 cells with more than 200 counts, less than 10 times the median counts, and less than 80% mitochondrial reads. We then obtained a UMAP of this data and transferred the mouse labels using the Seurat R package (Supplemental Fig. S15)

Assessment of the accuracy of deconvolution

One classical first check of deconvolutions is to verify that the proportions of cell types estimated in the single cell dataset are recovered in a pseudobulk dataset. Pseudobulk datasets are artificial datasets obtained by adding up the counts for all cells of the same sample. We then correlated these estimated proportions to the real proportions (real cell counts from the scRNA-seq data). We built pseudobulks for every sample (individual) and showed that the recovered proportions are well correlated to cell counts for the rat scRNA-seq reference (Supplemental Fig. S15). When we performed a similar experiment with the mouse scRNA-seq data, the recovered proportions were much less well correlated (Supplemental Fig. S15). Together with the high mitochondrial content and the high variability of cluster content between samples, this made us choose the rat dataset as a reference for deconvolution (Fig. 4). We decided to keep mouse and hamster scRNA-seq data for checking the marker conservation (Supplemental Fig. S15 B,E,H).

This verification using pseudobulk is a necessary step, but remains artificial since it is based on data generated using the same technology. We then represented cell proportions deconvolved using bulk data for the 3 *Rattus norvegicus* and 3 *Mus musculus* samples. We do not have the ground truth for the cell proportions in these individuals, but they should not differ much between mouse and rat according to literature (Clark et al. 2019). The results are represented on the Supplemental Fig. S16. This shows that cell proportions are not very similar between single cell (counts and estimated proportions) and bulk data (mus and rat). Some cell types are absent (Podo), some are vastly underestimated (DCT/CNT/PC, IC). Some others seem better recovered. In any case, we observed that the proportions do not differ between mouse and rat. Depending on the technology transfer (scRNA-seq and bulk RNA-seq do not capture the same portion of mRNA) and depending on the nature of cell types (biases due to cell fragility, cell size or natural doublets that are not captured), scRNAseq may not represent the real proportion of cells in native organ. Because these biases apply consistently to mouse and rat, and rat is at the most basal species of our Murinae dataset, we are confident regarding applicability of deconvolutions to the Murinae dataset.

To understand why some cell types are well estimated and some are not, we defined the top 100 discriminating marker genes for each cell type using Seurat *FindMarker* function (Supplemental Fig. S17).

- For Proximal Tubule (PT), the markers are very discriminating (that is, very highly expressed in this cell type, and weakly expressed elsewhere), and this cell type represents more than 50% of kidney cell types in histology measurements (Clark et al. 2019). It is overestimated by deconvolution, possibly because its markers benefit from non-linear effects in the PCRs during bulk library preparation.
- LOH is the second most common cell type (also in histology); LOH-associated markers are well distinguished from PT but not from DCT, probably leading to LOH underestimation by deconvolution.
- Endo and stroma are less abundant. The markers are not very specific between these cell types, but they discriminate them well from other cell types. Endo and Stroma proportions are recovered by deconvolutions, but they are interconnected.
- IC and DTC/CNT/PC are missing in deconvolutions, probably because their markers are not discriminating. They are expected in low proportion: 7% to 12%.
- Since podocytes are a rare cell type in histology (2–3%) and in the scRNA-seq data, they will be very difficult to detect. Furthermore, the marker genes are not perfect.

Altogether, this suggests that cell proportion in bulk data and the cell type specificity determines the success of deconvolution in a given cell type.

We do not see any reason why the difference in proportions we observe between arid and mesic environments would be created by an artifact of deconvolution. When we compared cell proportions obtained for arid and mesic conditions (Fig. 4), we observed that endo and stroma increase both in arid condition, to the detriment of the major cell types PT and LOH. This effect cannot be due to the interconnection of the proportions of endo and stroma. IC and

DTC/CNT/PC become detectable. Altogether, besides the limits of deconvolutions explained above, this suggests the reduction of PT in arid species at the benefit of other cell types.

Comparison of marker genes expression in 3 species

Average expression per cell type was obtained for each species by using the averageExpression function from Seurat. This was used with custom scripts to plot the heatmaps showing expression levels per cell type for the 179 marker genes described above (Supplemental Table S9) together with their associated cell type and their differentially expressed status in the Murinae dataset (Supplemental Fig. S12).

We observed that the marker genes tend to be expressed in similar cell types across species, including the hamster which is a Cricetidae (and not a *Murinae* like mouse and rat). We decided therefore to apply deconvolution on all species with expression data (Total_expr dataset).

Controls and limits in the interpretation of changes in cell proportions

Using rat reference, we observed convergent changes in proportions for several cell types (Fig 3). The convergent changes in proportions are consistent with convergent changes in many marker genes. This is for instance the case for the internal medullary collecting duct (CD), a cell type that selectively expresses *Aqp2* (Chen et al. 2017; Habuka et al. 2014; Miao et al. 2021). We found that arid *Murinae* species express *Aqp2* at a significantly higher level in bulk RNA-seq data and, accordingly, CD is found in a higher relative proportion. Conversely, *Slc28a1*, a marker gene of the proximal tubule (PT), is downregulated in arid *Murinae*, in accordance with a smaller number of PT cells inferred in these species.

The Murinae_expr dataset comprises mainly our own expression data (12/14 species), made on purpose for these deconvolutions. In consequence, we could control and homogenize the dissection process, and we extracted RNA from the whole kidney, avoiding using only parts of the kidney, which could bias the cell proportions. In the total_expr dataset, we used more RNA-seq data that we did not generate ourselves, for which we had no information on dissection and RNA preparation. Therefore, for these species we have more uncertainty on the validity of their position (or the position of replicates) on the PCA.

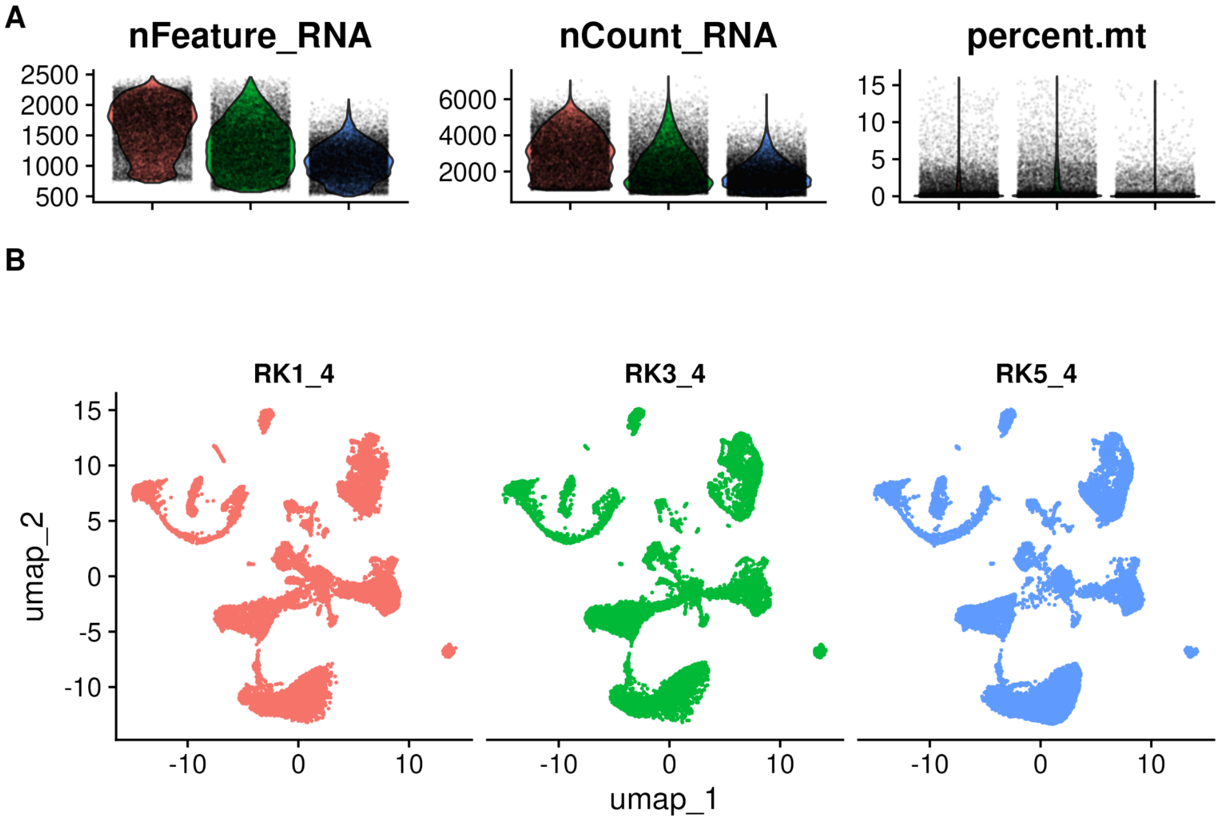
We interpret the convergence signal as a convergence on cell type proportions. In species with high capacity for urine concentration, the relative number of short loops is increased (Pannabecker 2013). Unfortunately, we lack resolution in the granularity of cell type annotations, particularly between PT segments (Chrysopoulou and Rinschen 2024) and between short and long loops of Henle to test this hypothesis.

Another possibility is that cell type identity may shift along the loops, which could also cause convergence in the deconvolution signal, even though it is not strictly a change in cell type proportions. For instance, as compared to rats, *Aqp1* was found to be expressed in a greater territory of the descending thin limbs of the loops of Henle in the kangaroo rats, which may allow greater solute concentration (Urity et al. 2012). Again, in our deconvolutions, we do not have the precision necessary to test this hypothesis.

Supplemental Tables and Figures

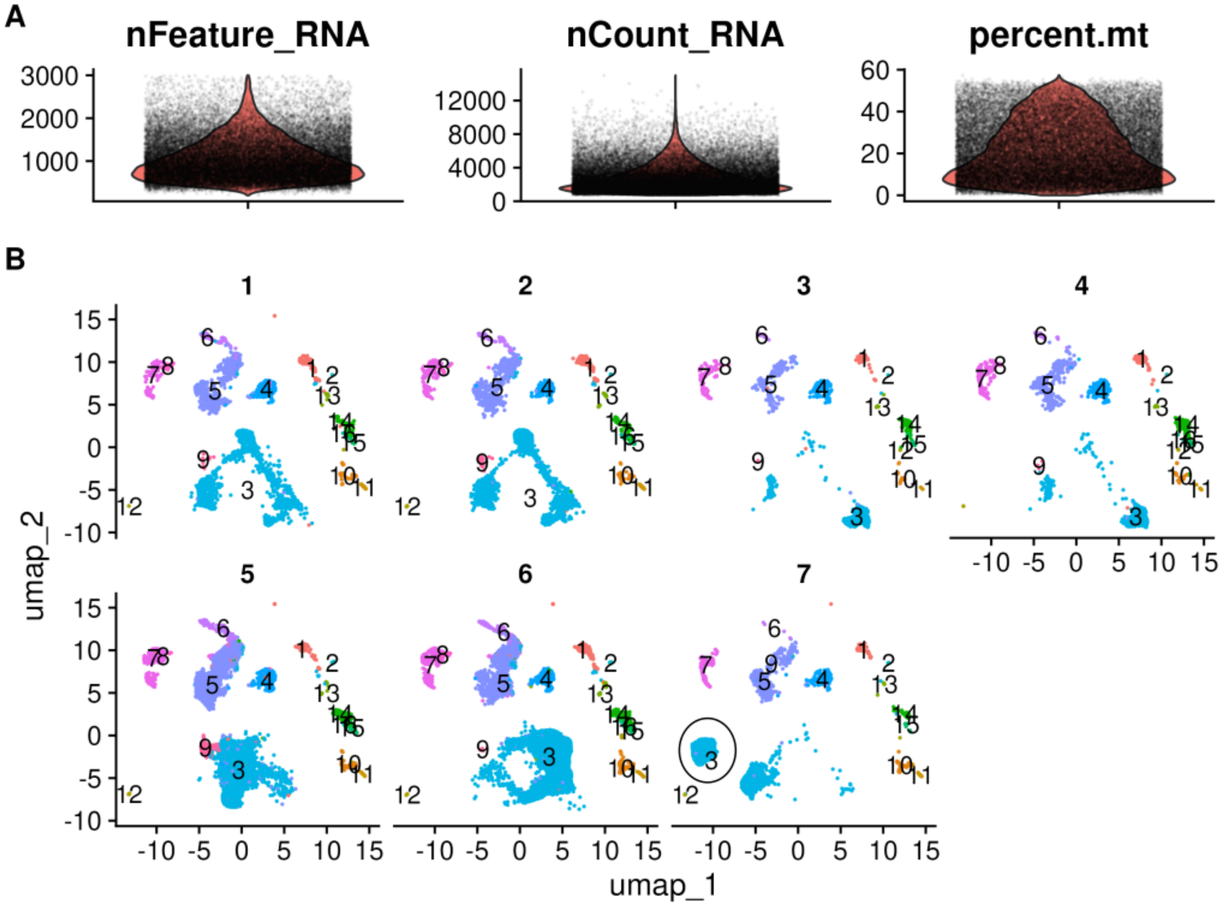
Supplemental Table S12: Cell proportions for all species.

Estimated proportion of cell types using MuSiC deconvolution tool, in the Total_expr dataset. Estimated proportions are given for all samples of the dataset.



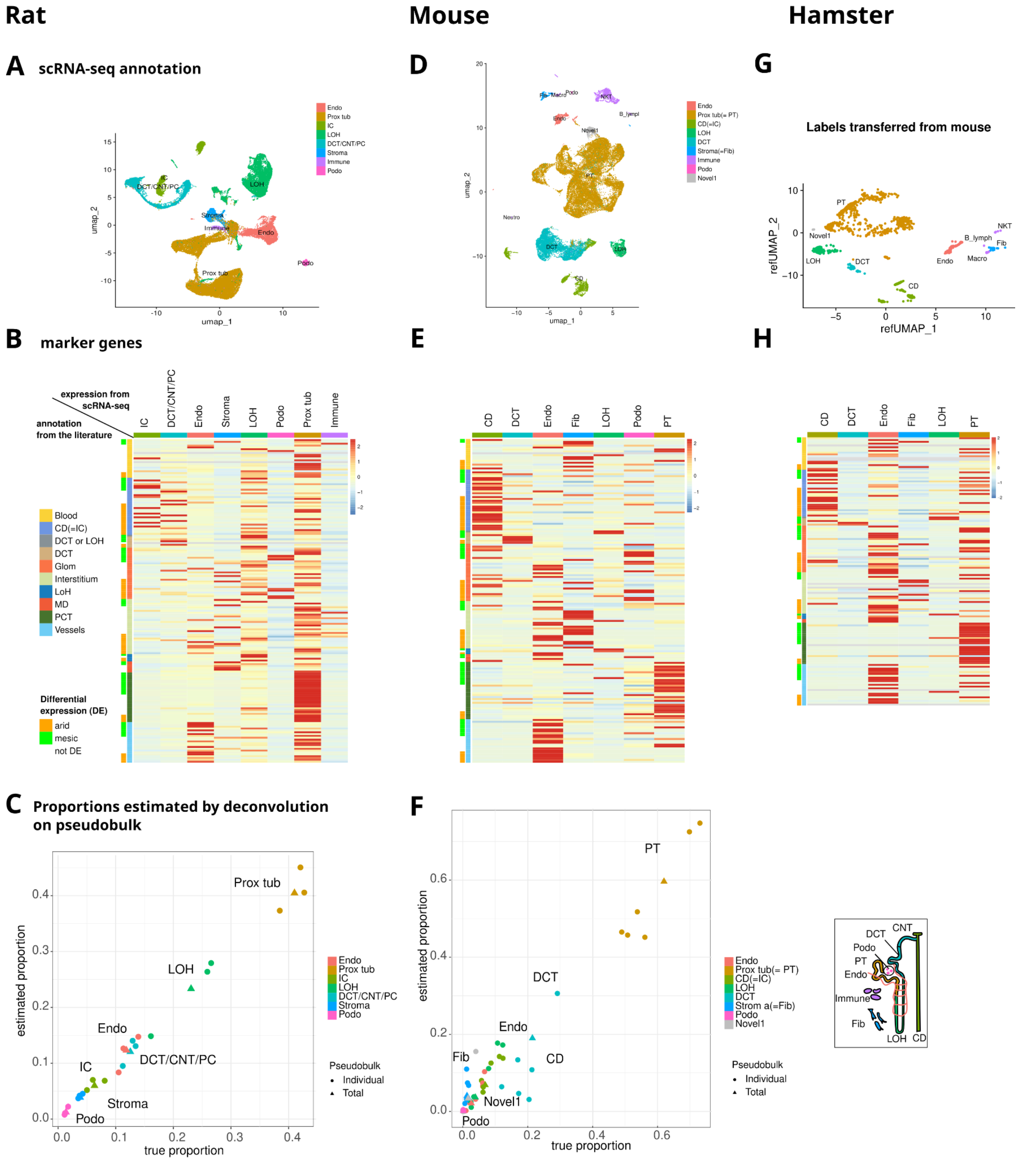
Supplemental Figure S13: Quality control for rat scRNA-seq data

The rat scRNA-seq data is taken from(Balzer et al. 2023). A- Violin plots representing the number of genes per cell, the number of reads per cell, and the percentage of mitochondrial reads per cell. B- UMAP of the data split per sample. The 3 samples were very homogeneous. Related to Fig. 4



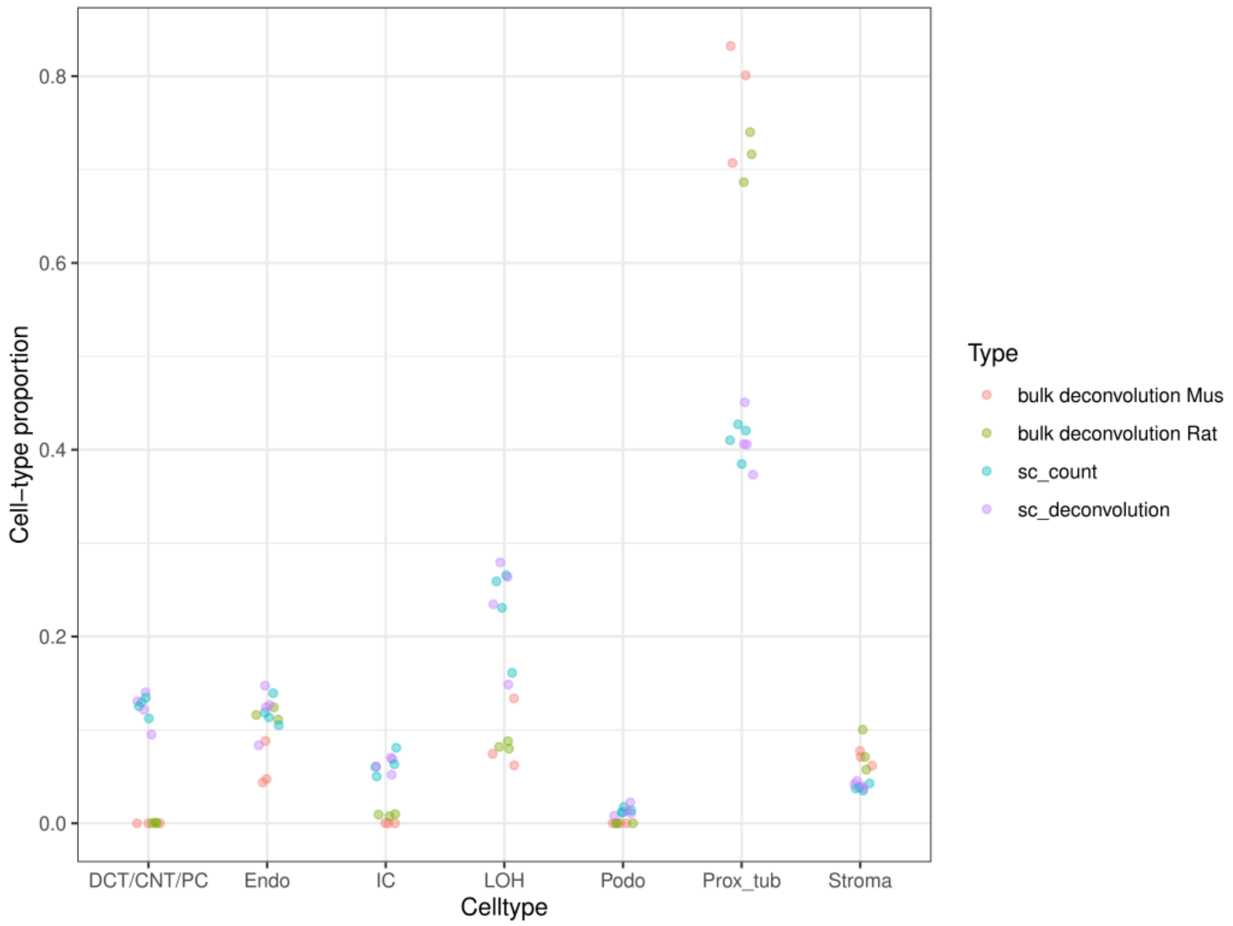
Supplemental Figure S14: Quality control for the mouse scRNA-seq data

The data is taken from (Park et al. 2018). A- Violin plots representing the number of genes per cell, the number of reads per cell, and the percentage of mitochondrial reads per cell. B- UMAP of the data split per sample. Cluster numbers from the parent study. ind 7 created an additional cluster (circle) and lacked several clusters in the published parent study (clusters 8,9). Related to Fig. 4


Rat
Mouse
Hamster
A scRNA-seq annotation
Endo
Prox tub
IC
LOH
DCT/CNT/PC
Stroma
Immune
Podo
umap_1
umap_2
Prox tub
LOH
Podo
D
20
10
0
−10
−10
−5
0
5
umap_1
umap_2
Fib
Macro
Podo
NKT
B_lympl
Endo
Novel1
PT
Neutro
DCT
LOH
CD
Prox tub(= PT)
CD(=IC)
Stroma(=Fib)
Novel1
G
Labels transferred from mouse
refUMAP_1
refUMAP_2
PT
Novel1
LOH
DCT
CD
Endo
B_lymph
NKT
Fib
Macro
B marker genes
expression from scRNA-seq
annotation from the literature
Blood
CD(=IC)
DCT or LOH
DCT
Glom
Interstitium
LoH
MD
PCT
Vessels
Differential expression (DE)
arid
mesic
not DE
E
H
C Proportions estimated by deconvolution on pseudobulk
estimated proportion
true proportion
Pseudobulk
Individual
Total
F
Strom a(=Fib)
DCT
CNT
Podo
PT
Endo
Immune
Fib
LOH
CD

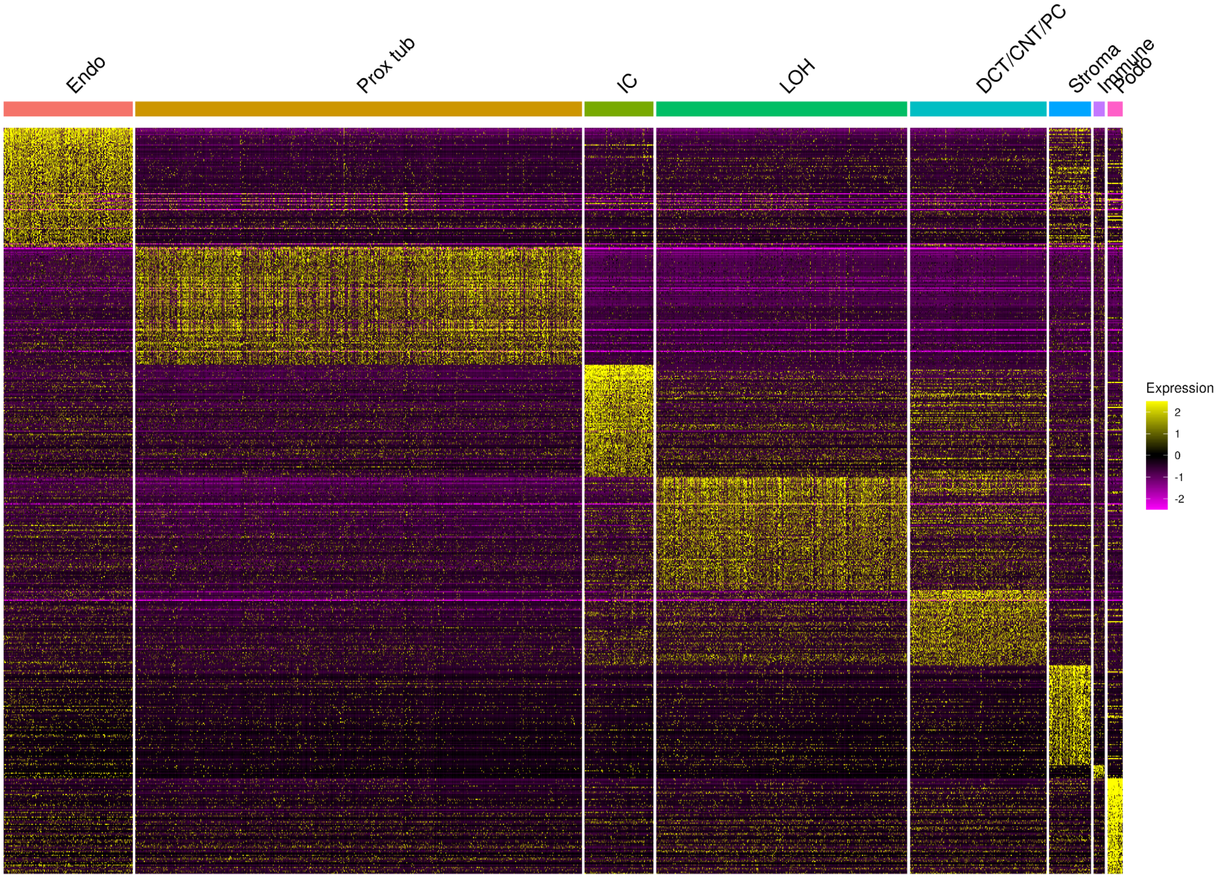
Supplemental Figure S15: Benchmark of the scRNA-seq reference

Kidney single cell RNA-seq data from rat, mouse and hamster with annotated cell types and known marker genes. UMAPs of the single cell data are represented for rat (A, 48,744 cells), mouse (D, 38,403 cells) and hamster (G, 1,924 cells, annotation transferred from mouse, no podocytes were recovered in hamster). Cell types have slightly different names in different species because the names were taken from the original publications, but they are shown in the same order and with matching colors. Cell types are abbreviated as follows in rat/mouse: collecting duct and intercalated cells: IC/CD. Distal convoluted tubule DCT|CNT|PC / DCT. Endothelium: Endo. Loop of Henle: LOH. Podocyte: Podo. Proximal Tubule: Prox Tub/PT. Immune cells (multiple types in mouse data, not represented in E for sake of place). (B,E,H) Marker gene expression levels taken from the average levels of expression per cell type are represented in each species. The annotations taken from the literature are presented in the color bar on the left. Genes differentially expressed (DE) between arid and mesic environments are indicated by orange (upregulated in arid) and green (upregulated in mesic) colors. The number of markers found expressed in scRNA-seq data was 179 (mouse and rat) and 154 (hamster) respectively. C) Cell type proportions in the rat scRNA-seq data, observed and estimated by deconvolution. Cell proportions are shown per sample (circles, 3 individuals) and for a pseudobulk aggregating the full dataset with all 3 samples (triangles). F) Same as C for the mouse scRNA-seq data. The recovery of cell proportions is much better with rat scRNA-seq (C). Related to Fig. 4



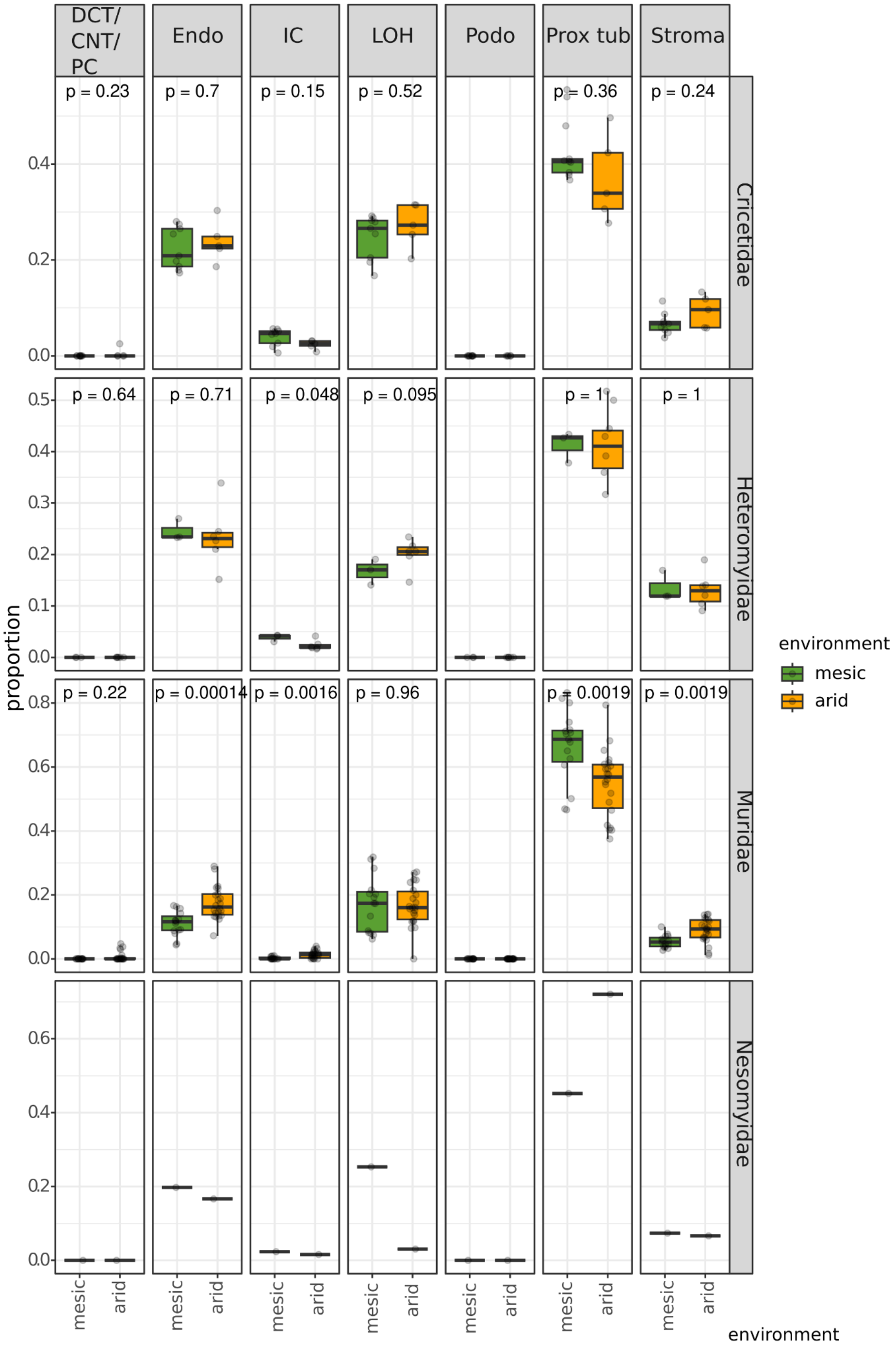
Supplemental Figure S16. Cell type proportions in mouse and rat

Cell type proportions obtained by deconvolution of bulk data in mouse and rat (3 samples in each species), by using single cell counts in the rat scRNA-seq reference, and by single cell deconvolution of pseudobulk of the rat reference. Related to Fig. 4.



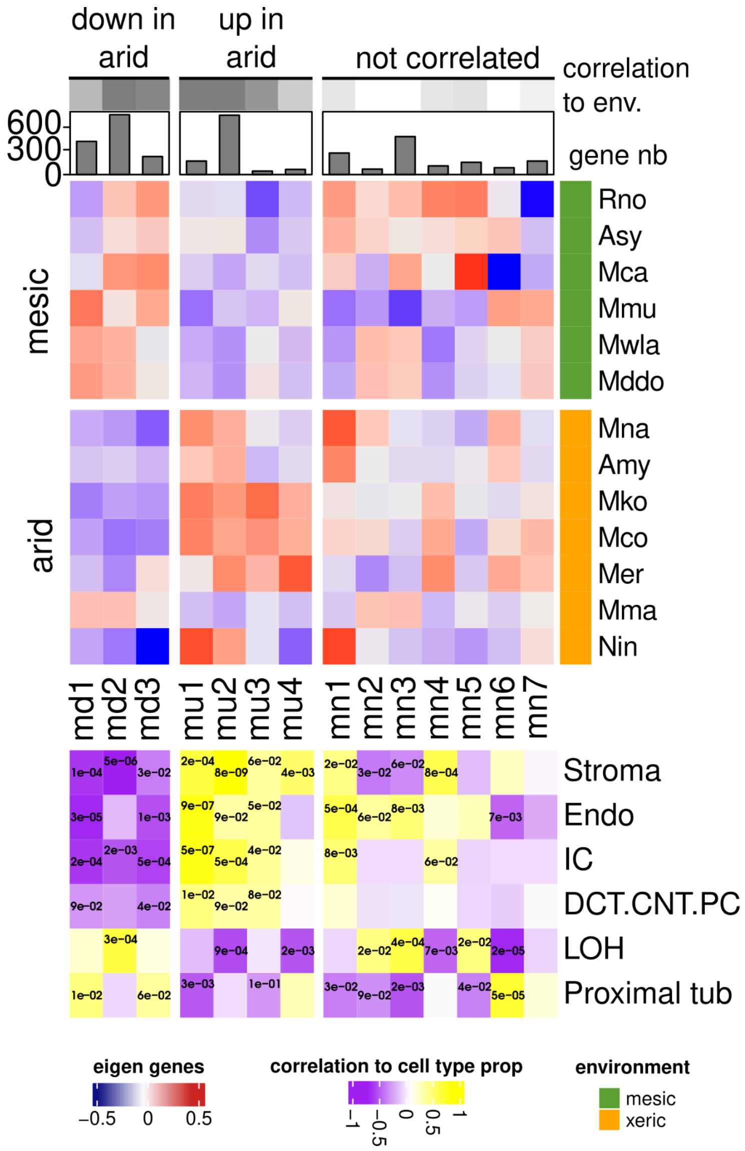
Supplemental Figure S17 Top 100 marker genes in the rat scRNA-seq data, for each cell type.

The Heatmap represents genes in rows and cells in columns. Gene relative expression levels are indicated by a color code. Related to Fig. 4.



Supplemental Figure S18: Cell proportions by family.

Cell proportions estimated by deconvolution using rat single cell RNA-seq data. Four rodent families which contain at least one species living in the arid and one living in the mesic environment are represented. Each dot is a single individual. P-values for Wilcoxon tests contrasting mesic and arid values by family are represented. Related to Fig. 4.



Supplemental Figure S19: Correlations of Murinae co-expression modules with cell proportions.

Top (As in Fig 4F): Co-expression modules are represented by their eigen genes with colors representing expression levels in each species. Modules significantly down or upregulated in arid species are depicted at the top of the panel, with grayscale colors indicating the strength of the correlation. Modules are named according to the sign and significance of this correlation and a number (mu: Murinae up; md: Murinae down; mn: Murinae not correlated). Barplots represent the numbers of genes in each module. Bottom: Heatmap showing the correlations between species expression levels (from eigen genes of WGCNA modules) and cell proportions (from deconvolutions). Correlations are color-coded and p values are indicated if significant at the 0.05 threshold. The colors highlight different cell types in mu and md modules. Related to Fig. 3 and Fig. 4.

Appendix 4. Patterns of convergent evolution in coding sequences

Supplemental Methods and results:

Supplemental Methods for Alignments/Species tree and Pelican analyses

To infer the species tree we obtained Nucleotides (NT) alignments for the complete families (with no missing sequence for any species of the total dataset and sites with no indel). NT alignments were obtained by backtranslating AA alignments using a custom script (https://gitbio.ens-lyon.fr/LBMC/cigogne/convergent_aridity_2024/-/blob/main/dataset_building_ropipe/scripts/back_translate_ali.py).

This script takes as input the aligned amino acid fasta file and the unaligned nucleotide sequences (CDS) files. Then, for each sequence, it goes through both sequences (AA and NT) in parallel and introduces indels into the nucleotide sequence when the aligned amino acid sequence contains indels. At the end, there is a checking step where the translation of the final aligned nucleotide sequence is compared to the initial amino acid sequence. We then used these NT alignments for phylogenetic reconstruction as explained in main methods. Alignments are available in Supplemental Data S6.

Pelican was used on the Total_seq and Murinae_seq datasets independently. Pelican uses as input the cleaned amino-acid (AA) alignments to detect convergent changes, as well as a species tree inferred as explained above. The tree was annotated with environmental conditions at leaves and internal branches using Supplemental Fig. S1. We applied the multinomial filter to speed up the analysis. We used a conservative threshold (multinomial-filter = 0.8) to remove sites that were sure not to yield a low p-value with Pelican. Further, we analyzed only sites that had substitutions in at least two clades, by using a custom script provided in Supplementary codes.

We have not investigated how sequence alignment affects our conclusions, but observation by eye of the sequence alignments led us to trust the alignments. At the scale of rodents, divergence is limited and alignments do not appear to be very difficult. Nonetheless, sequence alignment remains a critical part of a phylogenetic pipeline. It is possible that better sequence alignments would produce slightly improved results, leading to more significant functional enrichments. Alternatively, it seems unlikely that the functional enrichments we do observe (e.g., SLC transporters) are due to alignment errors, because we do not see why alignment errors would affect SLC genes more than other genes. We believe our search for convergent sequence evolution in response to adaptation to arid environments would most benefit from an increased number of species, which is outside of the scope of this study.

We set up a threshold for Pelican p-values that controls for the rate of false positives, by using negative control simulations (see below). We also used positive control simulations to evaluate the expected p-values for true positive sites of different kinds.

Positive control simulations.

For each gene, we selected the best site as detected by Pelican as the one with the lowest p-value. We then obtained for this site the main amino-acid in mesic species by selecting the amino-acid with the largest proportion in these species, and the main amino-acid in arid species.

- For the murinae_seq data: For each distinct pair of amino-acids (161) we simulated all combinations of 1 to 5 transitions to the arid environment (380). We obtained 61,180 simulated sites for which we scored convergent evolution with Pelican.
- For the total_seq data: We used the same process as in murinae_seq data. For each distinct pair of amino-acids in the total_seq dataset (131) we simulated combinations of 1 to 22 transitions to the arid environment. For each pair of amino-acids, as there were too many alternative combinations with up to 22 transitions to an arid environment (e.g., there are P(22,5)=2.44188 E+6 ways to place 5 transitions out of 22), we chose 200 combinations of transitions at random. We obtained 477,233 simulated sites for which we scored convergent evolution with Pelican. We represent only the sites with 1 to 6 transitions in the Supplemental Fig. S20 as they were used for defining the threshold (Fig 5B).

We obtained a distribution of p-values for each scenario. This distribution is represented Fig 5B for murinae_seq and Supplemental Fig. S20 for total_seq.

Negative control simulations:

We used the simulation tool Pastek (https://gitlab.in2p3.fr/pveber/pastek, commit: d267be5f) to simulate 100000 sites, using murinae_seq and total_seq tree topologies without annotations. We used the profiles where all transitions are equally likely to simulate random evolution. We kept only non-constant sites and scored the convergence with Pelican. We obtain a distribution of p-values and put the threshold so as to accept 1/1000 of these negative controls. This gives us the false positive rate expected by random evolution unrelated to the arid/mesic status. This distribution is represented in Fig 5B for murinae_seq dataset and Supplemental Fig. S20 for total_seq dataset.

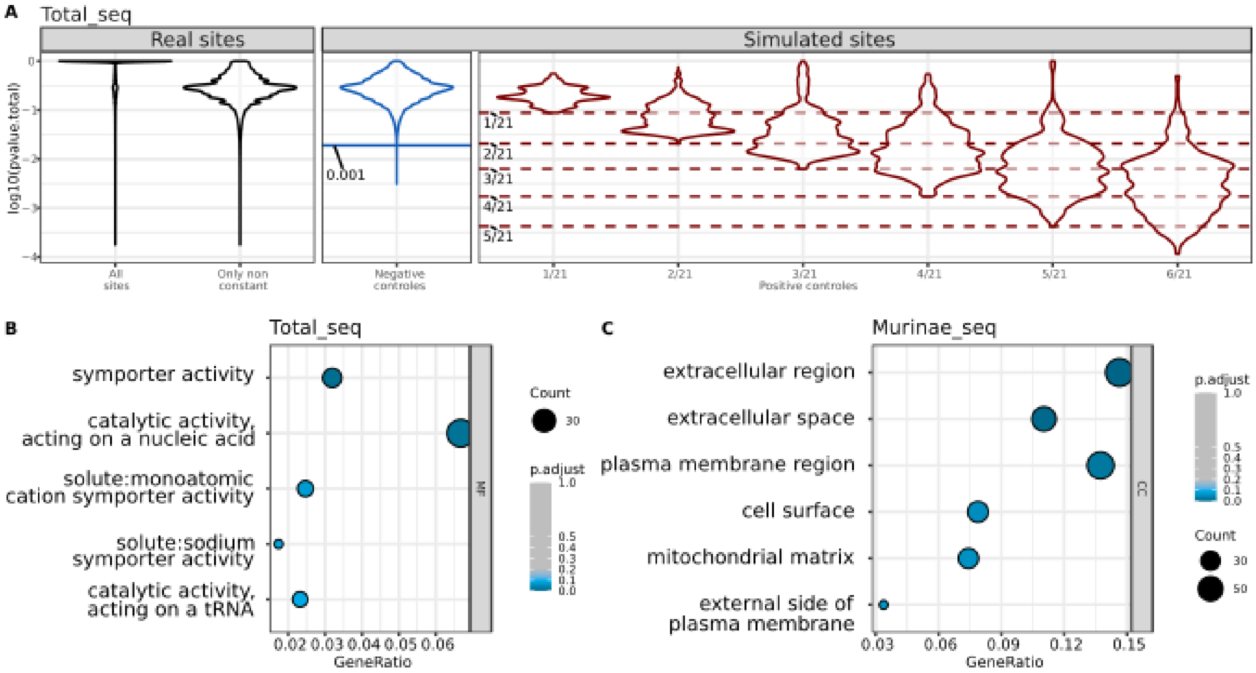
We selected genes with at least one site with a p-value below the threshold and performed an overrepresentation analysis using the Gene Ontology database (ClusterProfiler (Wu et al. 2021)

Supplemental Data

Supplemental Data S6: Multi-species alignments in Murinae_seq and Total_seq datasets

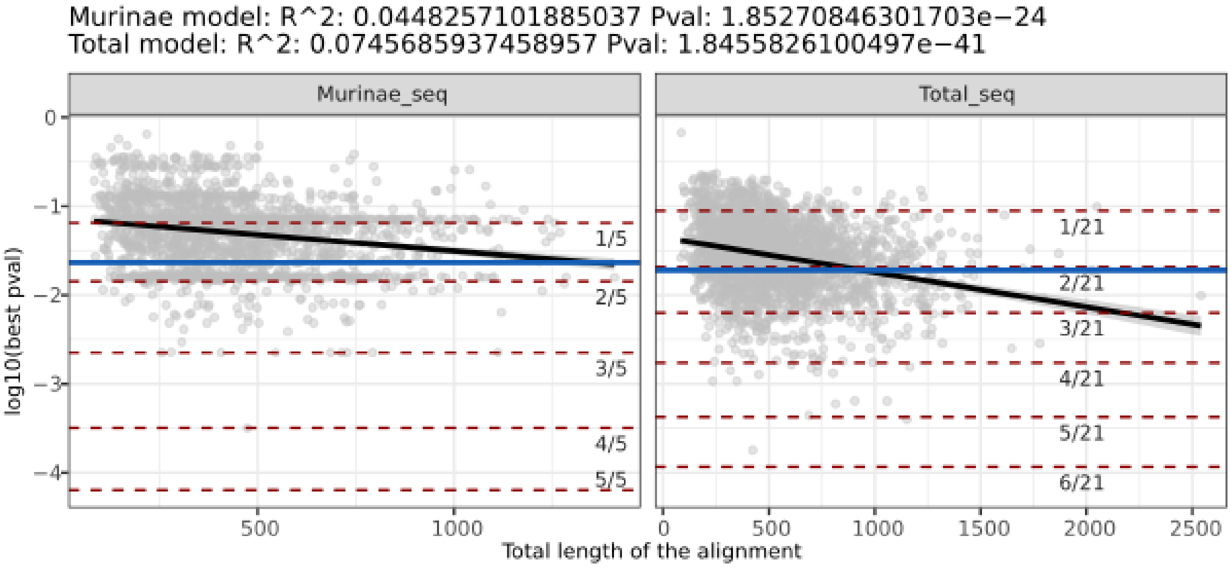
Supplemental Data S7: Pelican results for all Murinae_seq and Total_seq, positive and negative control simulations.

Supplemental Figures



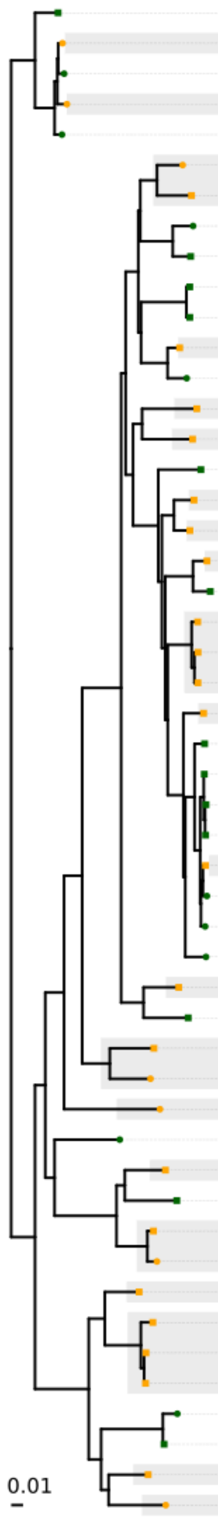
Supplemental Figure S20: Pelican threshold selection for Total_seq dataset and enrichments for Total-seq and Murinae-seq

(A) Distributions of $\log_{10}$(p value) of pelican tests for real data (black; all sites and excluding constant sites), simulated negative controls (blue: 1/1000 quantile is indicated), simulated positive controls (red: number of transitions indicated; dashed lines show the minimum p-value). (B) Overrepresented Gene Ontology terms found in the Total_seq dataset by comparing genes with at least one significant site with Pelican, to the rest of tested genes (adjusted p-value <0.1). (C) Same for the Murinae_seq dataset. Related to Fig. 5.



Supplemental Figure S21: Pelican p-value and alignment length

Relationship between the lowest p-value of each gene and the length of the alignment, for the Murinae_seq and the Total_seq datasets. Threshold of detection (in blue) and positive controls (in red dashed lines with number of transitions) are indicated. The squared correlations are indicated on top. Related to Fig. 5.



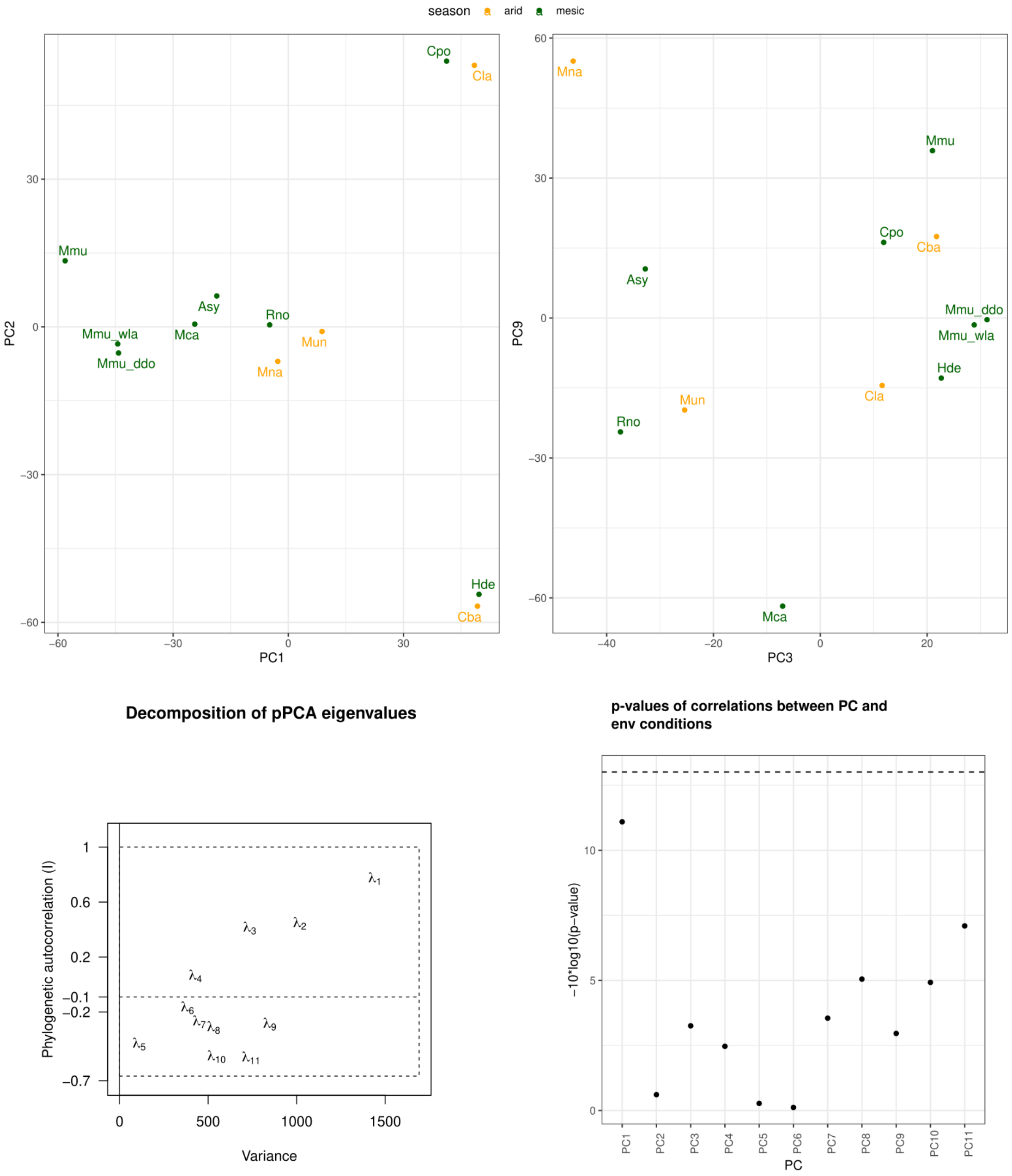
	Ubr7_224	Jak1_143	Pja2_303	Tut1_759	Atp13a2_428	Lig4_647	Tspan8_158	Dnase2a_110	Nop53_190	Xylt2_721	Slc15a2_148	Pkdcc_230	Immt_258	Vars_638	Ece1_325	Atg16l1_350	Syde2_203	Tpk1_121	Slc4a1_47	Pip5k1a_318
Tamias sibericus	I	V	I	N	A	I	A	S	R	G	V	S	I	I	S	A	S	M	K	N
Urocitellus_parryii	I	V	I	N	A	V	P	S	R	G	V	S	V	V		A	S	V	K	N
Ictidomys_tridecemlineatus	I	V	I	N	A	V	A	S	R	G	V	S	V	V	S	A	S	V	K	N
Spermophilus_dauricus	I	V	V	N		V	A	S	R	G	V	S	V		S	A	S	V	K	N
Marmota_marmota	I	V	I	N	A	V	A	S		G	I	S	V	I	S	A	S	V	K	N
Cricetulus_griseus	V	I	V	D	I	I	A	S		G	L		V	V	S	V	R	V	R	N
Mesocricetus_auratus	V	V	V	D	I	I		S	R	G		G	V	V	–	V	P		R	
Microtus_ochrogaster	I	V	I	N	V	V	A	S	R	G		S	V	I	S	A	S	M	K	N
Myodes_glareolus	I	V	I	N	V	V	A	S	R	G	I	S	V	I	S	A	S	M	K	N
Abrothrix_olivaceus	F	V	I	N	V	V	A	S	R	G	I	S	V	I	S	A	S	M	K	S
Abrothrix_longipilis	F	V	I	N	V	V	A	S	R	G	I	S	V	I	S	A	S	M	K	S
Peromyscus_eremicus	I	V	I	N	V	V	A	S	R	G	V	S	V	I	S	V	S	M	K	N
Peromyscus_maniculatus	I	V	I	N	V	V	A	S	R	G	V	S	V	I	S	A	S	M	K	N
Acomys_dimidiatus	V	I	I	D	I	I	P	S	R	G	L	G	V	I	T	V	S	V	R	T
Meriones_unguiculatus	V	I	I	N	I		P	A	R	S	V	G	V	I	T	A	R	V	R	S
Rattus_norvegicus	I	V	I	K	I	I	A	S	R	G	V	S	I	I	S	A	S	V	K	S
Arvicanthis_niloticus	V	V	I	N	I	I	A	T	R	–	V	S	V	V	S	A	N	V	K	N
Micaelamys_namaquensis	I	V	I	N	I	V	A	S	Q	S	V	G	I	I	S	A		V	K	N
Apodemus_mystacinus	I	V	I	D	I	I	A	S	R	G	V	S	V	I	S	A	S	V	K	S
Apodemus_sylvaticus	I	V	I	N	V	I	A	S	R	–	V	S	I	I	S	A	S	M	K	S
Mastomys_kollmanspergeri	I	A	I	D	I	I	A	T	Q	G	V	G	V	I	S	V	S	V	K	S
Mastomys_erythroleucus	I	A	I	D	I	I	A	T	R	G	V	G	V	I	S	V	S	V	K	S
Mastomys_coucha	I	A	I	D	I	I	A	T	R	G	V	G	V	I	S	V	S	V	K	S
Nannomys_indutus	S		I	N	V		A	T	R		V	S	V		T			V	–	
Mus_caroli	S	V	I	N	V	V	A	S	R	G	V	S	I	I	S	A	S	M	K	N
Mus_musculus	S	V	I	N	V	V	A	S	R	G	V	S	I	I	S	A	S	–	K	N
Mus_musculus_wla	S	V	I	N	V	V	A	S	R	G	V	S	I	I	S	A	S	V	K	
Mus_musculus_ddo	S	V	I	N	V	V	A	S	R	G	V	S	I	I	S	A	S	V	K	N
Mus_macedonicus	S	V	I	N	V	V	A	S	R	G	V	S	I	I	S	A	S	V	K	N
Mus_spicilegus	S	V	I	N	V	V	–	S	R	G	V	S	I	I	S	A	S	V	K	N
Mus_spretus	S	V	I	N	V	V	A	S	R	G	V	S	I	I	S	A	S	V	K	N
Mus_pahari	S	V	I	N	V	V	A	S	R	G	V	S	V	I	S	A	S	M	K	N
Saccostomus_campestris	I	V	I	N	I	I	A	T	R	G	V	S	V	V	S	A	S	V	K	S
Dendromus_melanotis	–	V	I	N	V		A	S	R	G	V	S	L	V	S	A	S	M	K	N
Eospalax_fontanierii	T	V	V	D	I		A	T	Q	S	V	G	V	V	S	V	R	V	R	S
Nannospalax_galili	T	V	V	D	I	I	A	T	Q	S	V	G	V	I	S	V	R	L	R	S
Jaculus_jaculus	T	I	V	K	I	I	P	S	Q	S	L	S	V	V	T	V	G		R	S
Castor_canadensis	M	V	I		V	I	A	S	R	G		S	V	I	S	A	S		R	S
Chaetodipus_baileyi	V	I	I	K	I	I	P	S	R	G	V	A	V	V	S	A	N	V	K	S
Heteromys_desmarestianus	I	V	I	N	I	V	A	S	R	G	I	S	I	I	S	T	S	V	K	N
Dipodomys_spectabilis	I	V	I		I	V	A	S	R	G	V	S	V	V	S	A	S	V	K	S
Dipodomys_ordii	I	V	I	N	I	V	A	S		G	V	S		V	S	A	S	V	K	S
Heterocephalus_glaber	V	V	I	S	I	I	A	V	R	G	I	S	V	V	S	V	S	V	R	T
Fukomys_mechowii	V	A	V	D	M	I	A	A	R	S	I	S	V	V	S	V	T	V	R	S
Fukomys_micklemi	V	A	V			I	A				I	S	V		S	V	T	V		S
Fukomys_damarensis	V	A	V	D	M	I	A	A	R	S	I	S	V	V	S	V	T	V		S
Cavia_aperea		V	I	–	V	V	A	S	R	G	I	S	I	I	S	V	–	C	K	N
Cavia_porcellus	I	V	I	G	V	V	A	S	R	G	I	S	I	I	S	V	S	V	K	N
Chinchilla_lanigera	T	V	I	G	I	I	P	V	R	G	I	S	V	V	T	V	N	V	T	S
Octodon_degus	V	A	I	G	V	I		T	Q	G	L	S	V	V	T	V	S	V	–	S

Supplemental Figure S22: Pelican best sites for the Total_seq dataset

Top 20 best sites found by PELICAN with the Total_seq dataset. Related to Fig. 5.

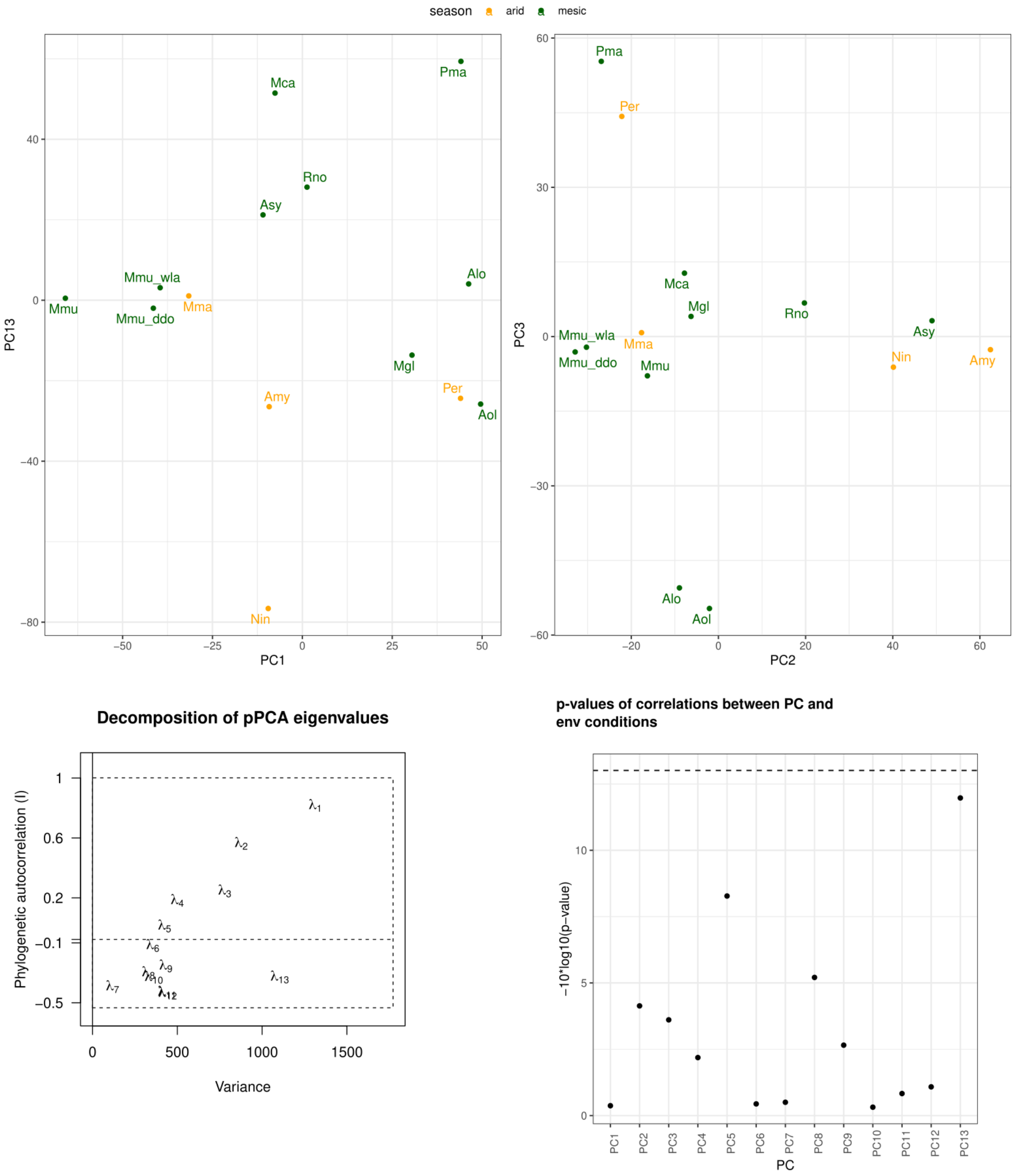
Appendix 5. Comparing ancient and recent adaptations to arid environment

Supplemental Figures



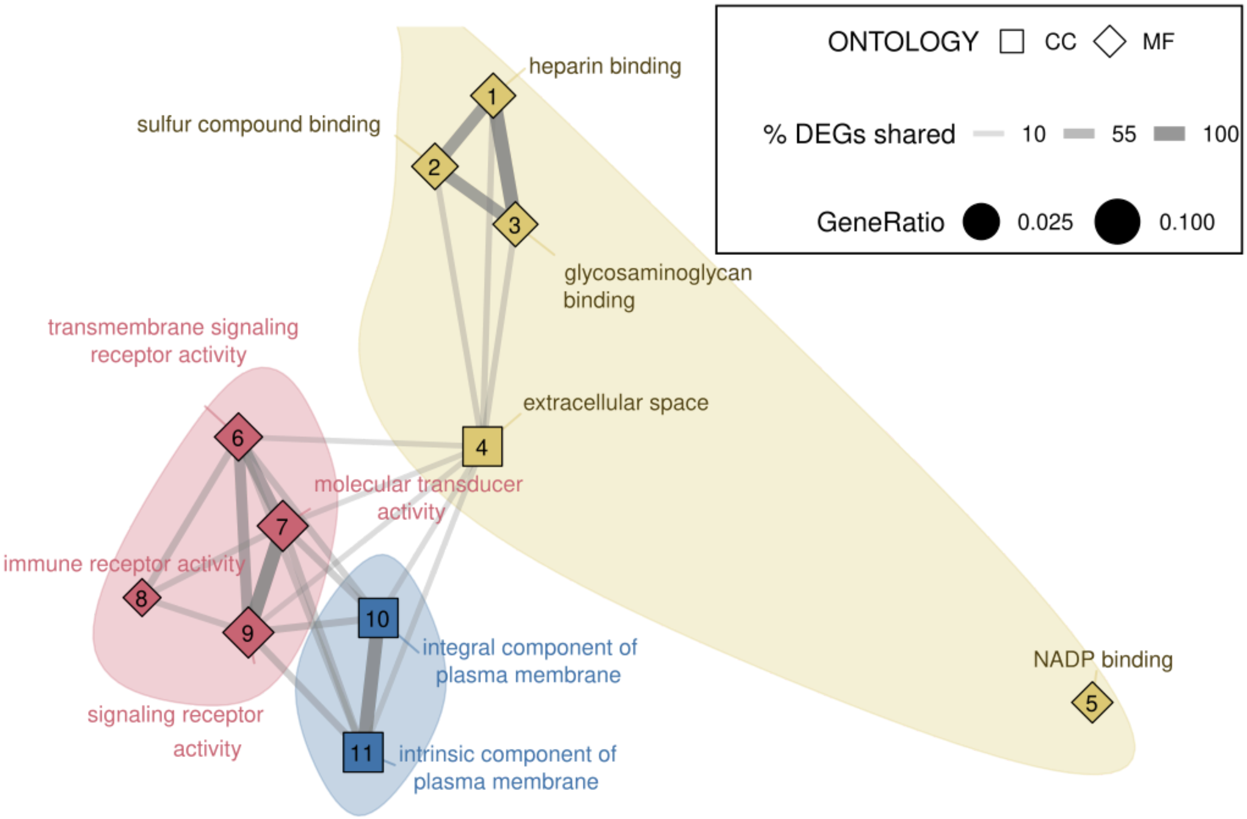
Supplemental Figure S23: phylogenetic PCA for the ancient dataset

Top: Species location on the map of the main components of the pPCA. Bottom: Decomposition of the pPCA eigenvalues showing positive phylogenetic autocorrelation (PC1, PC2, PC3) and negative phylogenetic autocorrelation (PC9). P-values of t.test between PCs and arid/mesic conditions, dashed line represents p=0.05 threshold. Related to Fig. 6.



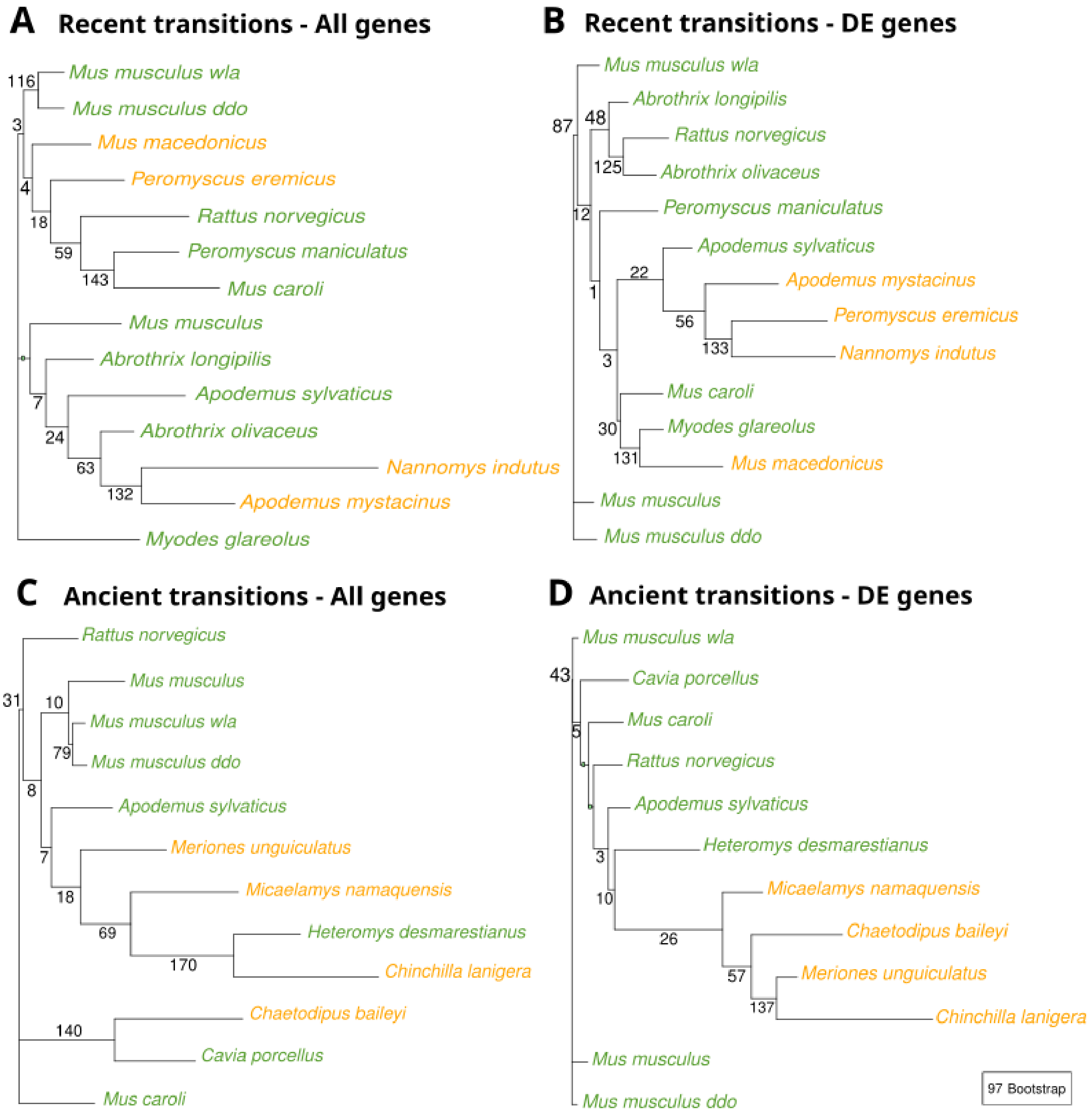
Supplemental Figure S24: phylogenetic PCA for the recent dataset

Top: Species location on the map of the main components of the pPCA. Bottom: Decomposition of the pPCA eigenvalues showing positive phylogenetic autocorrelation (PC1, PC2, PC3) and negative phylogenetic autocorrelation (PC13). P-values of t.test between PC coordinates and arid/mesic conditions, dashed line represents p=0.05 threshold. Related to Fig. 6.


ONTOLOGY □ CC ◇ MF
% DEGs shared 10 55 100
GeneRatio ● 0.025 ● 0.100
heparin binding
sulfur compound binding
glycosaminoglycan binding
transmembrane signaling receptor activity
extracellular space
molecular transducer activity
immune receptor activity
integral component of plasma membrane
NADP binding
signaling receptor activity
intrinsic component of plasma membrane
1
2
3
4
5
6
7
8
9
10
11

Supplemental Figure S25: Gene Ontology enrichment in differentially expressed genes from the "Ancient transitions" dataset.

Functional enrichments in the differentially expressed genes in the "Ancient transitions" dataset. All significantly enriched Biological process terms were used (FDR = 0.05), links on the graph represent the percentage of genes in common between two terms. Only edges representing at least 10% of shared genes are drawn. Term names are indicated on the graph. Related to Fig. 6.



A Recent transitions - All genes

B Recent transitions - DE genes

C Ancient transitions - All genes

D Ancient transitions - DE genes

Supplemental Figure S26. Dendrograms for recent and ancient datasets

Neighbor-joining reconstructed dendrograms of the Recent_exp and Ancient_exp datasets from expression data. We reconstructed dendrograms of the Recent_exp (A and B) and the Ancient_exp (C and D) datasets using either all the expressed genes or only differentially expressed (DE) genes. We used the (1-Spearman) distance matrix and reconstructed the neighbor joining tree using ape R package with 1000 bootstraps. Related to Fig. 6.

Appendix 6. Comparison of convergent genes between datasets and with the literature

Supplemental Results

Comparison of the sets of differentially expressed genes between datasets

We compared the genes with repeated changes in expression in different datasets to see whether the processes involved are the same. There was only one common gene between our 4 datasets: *Fam20a*, which is upregulated in arid species. *Fam20a* also shows some degree of convergent amino acid profiles (Fig. 5D and E). Furthermore, nine genes were shared by the Murinae_expr, the “ancient transitions” and the total_expr datasets: *Marc1*, *Casr*, *Lgals3*, *Ctsh*, *Pqlc3*, *Slc35b4*, *Ring1*, *Epb41l4a* and finally *Fam20a*.

Between the datasets with most differentially expressed genes, the Murinae_expr (744 genes) and the “ancient transitions” dataset (632 genes), we found an overlap of 90 for differentially expressed genes, of which 75 are biased in the same sense (24 up and 51 down).

Related to discussion: Gene functions and overlap with previous studies

Our expression comparisons revealed a significant amount of genes associated with kidney physiology. Among the common physiological systems allowing mammalian survival in deserts described in a recent survey, there were increased urine osmolarity and increased water reabsorption from the kidney, higher levels of plasma creatinine, increased plasma osmolality, change in insulin secretion for adaptive tolerance to dehydration and starvation (Rocha et al. 2021). We found in our data several genes and pathways relevant to these systems.

Aquaporins form a gene family of water transporters that has been associated with desert adaptation in rodents (Bittner, Mack, and Nachman 2022; Pannabecker 2015; Marra, Romero, and DeWoody 2014; Giorello et al. 2018). In the Murinae dataset, we found convergent upregulation of *Aqp2* and *Aqp7* in xeric species. *Aqp2* is the dominant water transport gene in the medullary Collecting Ducts. Since its spatial pattern of expression seems similar in many rodent species (Pannabecker 2013), we may have detected a change in intracellular expression level. Of note, because for some species we rely on *de novo* transcriptome assemblies, we cannot reconstruct the sequences of genes with very low levels of expression. *Aqp4* for instance, another important water transporter (Donald and Pannabecker 2015), is not available in our datasets, possibly for this reason. In a previous study, aquaporin expressions were shown to respond to hydric stress (MacManes 2017), but in our dataset we cannot discriminate between adaptation and plastic response.

We found that many solute carriers are differentially expressed (2 in the Total_expr: *Slc40a1* and *Slc35b4*, 17 in the Murinae_expr: *Slc25a33, Slc5a8, Slc14a2, Slc36a1, Slc27a2, Slc22a22, Slc35b41, Slc8b1, Slc43a3, Slc30a4, Slc39a5, Slc11a2, Slco4a1, Slc16a12, Slc16a10, Slc22a4, Slc25a35* and 18 in the ancient transition set : *Slc29a2, Slc39a51, Slc22a18, Slc2a5, Slc16a2, Slc2a12, Slc22a2, Slc14a1, Slco4a11, Slc30a1, Slc25a48, Slc6a6, Slc35b3, Slc44a3, Slc39a14, Slc35b42, Slc25a14, Slc26a1*). *Slc14a2*, which was upregulated in arid species in Murinae, is an urea transporter whose knock-out causes decreased urine osmolality (Fenton et al. 2004). *Slc8b1*, a calcium:sodium exchanger, was upregulated in arid species in Murinae and carried marks of positive selection in a previous study of adaptation to aridity in *Peromyscus* rodents (Tigano, Colella, and MacManes, n.d.).

We intersected our differentially expressed genes with results from a recent study of convergent adaptation to desert life in 3 pairs of rodent species (Bittner, Mack, and Nachman 2022). Among their list of genes with evidence for convergent differential expression and involvement in kidney physiology and/or signature of sequence selection, we found that one gene was differentially expressed in the Murinae dataset (*Fstl1*) and four were also differentially expressed in the "ancient transitions" dataset (*Col4a5, Pax2, Robo2* and *Bhlhe40*).

Appendix 7. Overview of data analyses workflows

Supplemental Method. Transcriptome assemblies

We removed adapters and low-quality bases (Q<20) using Trimmomatic version 0.38, with options "TRAILING:20 MINLEN:25 AVGQUAL:20" (Bolger, Lohse, and Usadel 2014). After this trimming, we checked the quality of the reads with FastQC. We then assembled the data with Trinity version 2.8.5 (Grabherr et al. 2011) with option "--full_cleanup". We predicted coding sequences from trinity assemblies with TransDecoder version 5.5.0, retaining only the best open reading frame per transcript, at least 80 amino-acids long (https://github.com/TransDecoder/TransDecoder). Basic quality values of assemblies, such as N50 and number of transcripts were retrieved with the implemented Trinity script trinityStats.pl (Haas et al. 2013). Completeness of gene repertoire was evaluated with BUSCO version 3.0.2 (Haas et al. 2013; Manni et al. 2021) with the mammalian library (mammalia_odb9).

Supplemental Method. Code availability

All codes are available in a gitlab repository (https://gitbio.ens-lyon.fr/LBMC/cigogne/convergent_aridity_2024) and in Supplemental Codes.

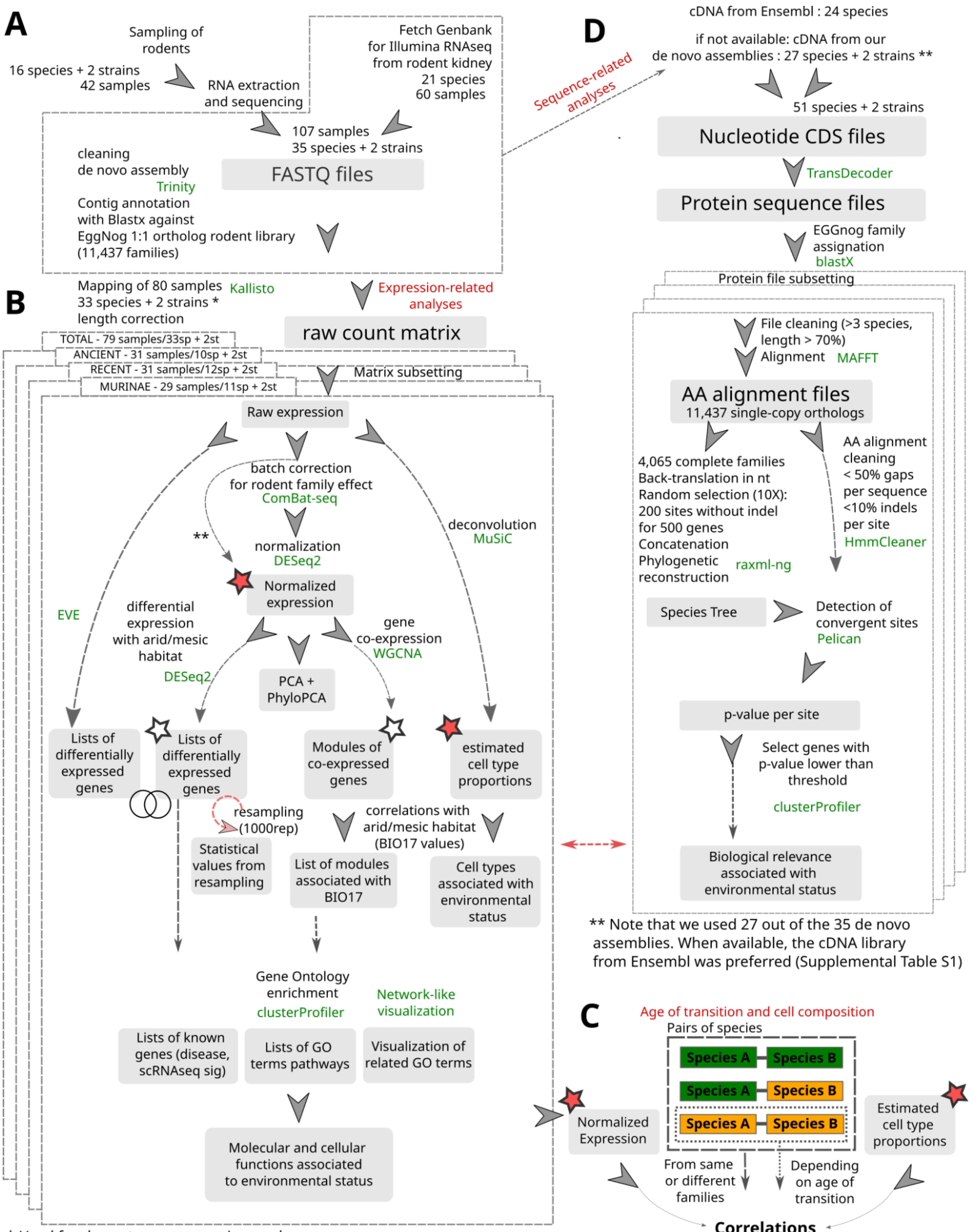
Supplemental Table and Figure

Supplemental Table S13: Characteristics of the *de novo* transcriptome assemblies.

Nb: Number of reads (millions); Med: median read size (base pairs); SE/PE: single or paired-end. Nb genes: Number of Trinity genes; Nb tr: Number of Trinity transcripts; mean and med tr size: mean and median transcript size; %full, %frag, % mis: % genes from BUSCO that are full, fragmented or missing respectively.

Species	**nb**	**Med**	**SE/ PE**	**Nb genes**	**Nb tr**	**N50**	**mean tr size**	**med tr size**	**% full**	**% frag**	**% mis**
Abrothrix longipilis hirtus	402.6	84	PE	117249	161868	3663	1343	432	86.5%	4.4%	9.1%
Abrothrix olivaceus	363	84	PE	192773	255533	3179	1152.2	403	87.9%	4.6%	7.5%
Cavia porcellus	101.2	50	PE	66483	86912	2594	1158.2	460	71.6%	9.8%	18.6%
Chaetodipus baileyi	168	100	PE	98886	128327	2743	1137.7	428	84.0%	5.5%	10.5%
Chinchilla lanigera	77	100	PE	159764	197049	2515	980.4	377	82,2%	5.8%	12.0%
Dipodomys spectabilis	32.6	100	PE	58199	75400	2140	1052.6	477	69.1%	11.6%	19.3%
Eospalax fontanierii baileyi	43.6	85	PE	41912	54610	2514	1296.0	630	70.7%	11.4%	17.9%
Fukomys damarensis	139.6	90	PE	232229	295210	1951	857.6	366	78.8%	10.2%	11.0%
Fukomys micklemi	41.8	200	PE	178563	274884	1482	897.7	488	75.5%	11.7%	12.8%
Heterocephalus glaber	84.1	50	SE	64456	67181	1575	816.8	391	65.2%	14.4%	20.4%
Heteromys desmarestianus	60.8	100	PE	67198	89416	2592	1182.9	482	78.4%	7.0%	14.6%
Meriones unguiculatus	120.2	50	PE	69152	88873	2690	1162.4	443	73.4%	7.6%	19.0%
Mesocricetus auratus	117.6	50	PE	93829	120135	2799	1111.1	394	78.3%	6.8%	14.9%
Mus caroli	107.2	100	PE	279491	310430	1300	672.3	321	81.1%	5.5%	13.4%
Myodes glareolus	82.4	125	PE	134576	181072	2545	1059.0	419	78.5%	7.4%	14.1%
Peromyscus eremicus	145.8	150	PE	247123	342093	1726	863.3	403	83.5%	5.8%	10.7%
Peromyscus maniculatus	75	150	PE	576943	859975	730	571.4	333	78.2%	8.1%	13.7%
Tamias sibiricus	80	50	PE	60647	75399	2083	1006.1	446	66.7%	12.1%	21.2%
Rattus norvegicus	98.8	50	PE	46384	57875	2568	1206.1	513	71.2%	8.7%	20.1%

Acomys dimidiatus	103.2	100	PE	134848	177587	2524	1069.6	433	80.7%	6.4%	12.9%
Apodemus mystacinus	123.6	100	PE	90667	124081	2462	1069.0	416	75.5%	9.0%	15.5%
Apodemus sylvaticus	121,8	100	PE	118005	162802	2641	1090.5	414	79.3%	6.4%	14.3%
Mastomys natalensis	129.4	100	PE	92974	126425	2289	1023.5	424	72.8%	10.6%	16.6%
Arvicanthis niloticus	111.2	100	PE	121749	163412	2706	1139.9	445	81.8%	6.0%	12.2%
Dendromus melanotis	131.6	100	PE	104615	142967	2584	1135.3	458	78.0%	8.8%	13.2%
Fukomys mechowii	128.6	100	PE	181943	243958	2801	1145.8	447	81.6%	8.0%	10.4%
Mastomys coucha	155	100	PE	99928	133785	2285	1026.8	430	75.3%	10.5%	14.2%
Mastomys erythroleucus	136.2	100	PE	130297	177804	2660	1116.8	440	82.4%	6.2%	11.4%
Mastomys kollmanspergeri	147	100	PE	97420	133683	2485	1078.6	428	77.3%	8.1%	14.6%
Micaelamys namaquensis	136.8	100	PE	75705	103369	2112	977.8	418	66.5%	12.9%	20.6%
Mus macedonicus	143	100	PE	135243	177997	2865	1157.9	443	81.9%	5.7%	12.4%
Mus musculus	104.8	100	PE	94222	126110	2747	1165.3	452	75.9%	7.3%	16.8%
Mus musculus ddo	153.6	100	PE	131029	171019	2715	1113.9	436	80.9%	5.6%	13.5%
Mus musculus wla	108.6	100	PE	102711	136796	2680	1120.6	432	78.1%	6.6%	15.3%
*Mus (Nannomys) sp.**	186.6	100	PE	150143	211669	2660	1096.5	418	80.0%	7.9%	12.1%
Mus (Nannomys) indutus	69.8	100	PE	52174	68374	1650	860.1	425	45.3%	21.7%	33.0%
Saccostomus campestris	119.2	100	PE	80959	109892	2319	1049.1	435	72.9%	10.6%	16.5%


A
Sampling of rodents
16 species + 2 strains
42 samples
RNA extraction and sequencing
Fetch Genbank for Illumina RNAseq from rodent kidney
21 species
60 samples
107 samples
35 species + 2 strains
FASTQ files
cleaning
de novo assembly
Trinity
Contig annotation with Blastx against EggNog 1:1 ortholog rodent library (11,437 families)
Sequence-related analyses
B
Mapping of 80 samples
33 species + 2 strains *
length correction
Kallisto
Expression-related analyses
raw count matrix
TOTAL - 79 samples/33sp + 2st
ANCIENT - 31 samples/10sp + 2st
RECENT - 31 samples/12sp + 2st
MURINAE - 29 samples/11sp + 2st
Matrix subsetting
Raw expression
batch correction for rodent family effect
ComBat-seq
**
normalization
DESeq2
deconvolution
MuSiC
Normalized expression
EVE
differential expression with arid/mesic habitat
DESeq2
gene co-expression
WGCNA
PCA + PhyloPCA
Lists of differentially expressed genes
Lists of differentially expressed genes
Modules of co-expressed genes
estimated cell type proportions
resampling (1000rep)
correlations with arid/mesic habitat (BIO17 values)
Statistical values from resampling
List of modules associated with BIO17
Cell types associated with environmental status
Gene Ontology enrichment
clusterProfiler
Network-like visualization
Lists of known genes (disease, scRNAseq sig)
Lists of GO terms pathways
Visualization of related GO terms
Molecular and cellular functions associated to environmental status
D
cDNA from Ensembl : 24 species
if not available: cDNA from our de novo assemblies : 27 species + 2 strains **
51 species + 2 strains
Nucleotide CDS files
TransDecoder
Protein sequence files
EGGnog family assignation
blastX
Protein file subsetting
File cleaning (>3 species, length > 70%)
Alignment
MAFFT
AA alignment files
11,437 single-copy orthologs
4,065 complete families
Back-translation in nt
Random selection (10X):
200 sites without indel for 500 genes
Concatenation
Phylogenetic reconstruction
raxml-ng
AA alignment cleaning
< 50% gaps per sequence
<10% indels per site
HmmCleaner
Species Tree
Detection of convergent sites
Pelican
p-value per site
Select genes with p-value lower than threshold
clusterProfiler
Biological relevance associated with environmental status
** Note that we used 27 out of the 35 de novo assemblies. When available, the cDNA library from Ensembl was preferred (Supplemental Table S1)
C
Age of transition and cell composition
Pairs of species
Species A
Species B
Species A
Species B
Species A
Species B
Normalized Expression
Estimated cell type proportions
From same or different families
Depending on age of transition
Correlations

* Used for downstream expression analyses

** no batch correction for the Murinae set

Supplemental Figure S27. General workflow to investigate convergent evolution at the level of whole transcriptomes, cell composition, gene expression and gene sequence.

Pipeline set up for the preparation of de novo assemblies (A), for downstream expression analyses (B and C) and for coding sequence analyses (D). (A) Main steps to generate de novo assemblies using RNA-seq data. FASTQ files were either produced for this study from lab/wild sampling or available online. (B) Expression-related analyses realized in the four datasets Total_seq, Murinae_seq, Ancient_seq and Recent_seq. Several output result files, including normalized count table, list of differentially expressed genes between arid and mesic species, list of modules containing co-expressed genes obtained from Weighted Gene Co-expression Network Analysis and estimated cell type proportion obtained from deconvolution analyses (all indicated by stars) are generated for enrichment and correlation analyses. (C) Normalized count table and estimated cell type proportion (red stars) were used to investigate the impact of the age of transitions. All possible pairs of mesic-mesic, mesic-arid or arid-arid species were defined, and correlations were performed using pairs from the same or different families. Further correlations were performed only on arid-arid pairs depending on the age of transitions. (D) Main computational steps to investigate changes in protein sequence alignments. Number of species and samples are indicated at the different steps of analyses. Note that not all samples used to generate *de novo* assemblies were used in subsequent expression analyses. Main softwares and R packages are indicated in green font.

Supplemental references

Aghová, Tatiana, Yuri Kimura, Josef Bryja, Gauthier Dobigny, Laurent Granjon, and Gael J. Kergoat. 2018. “Fossils Know It Best: Using a New Set of Fossil Calibrations to Improve the Temporal Phylogenetic Framework of Murid Rodents (Rodentia: Muridae).” *Molecular Phylogenetics and Evolution* 128 (November):98–111.

Aghová, T., K. Palupčíková, R. Šumbera, D. Frynta, L. A. Lavrenchenko, Y. Meheretu, J. Sádlová, et al. 2019. “Multiple Radiations of Spiny Mice (Rodentia: Acomys) in Dry Open Habitats of Afro-Arabia: Evidence from a Multi-Locus Phylogeny.” *BMC Evolutionary Biology* 19 (1): 69.

Avila Cobos, Francisco, José Alquicira-Hernandez, Joseph E. Powell, Pieter Mestdagh, and Katleen De Preter. 2020. “Benchmarking of Cell Type Deconvolution Pipelines for Transcriptomics Data.” *Nature Communications* 11 (1): 5650.

Balzer, Michael S., Mira Pavkovic, Julia Frederick, Amin Abedini, Alexius Freyberger, Julia Vienenkötter, Ilka Mathar, et al. 2023. “Treatment Effects of Soluble Guanylate Cyclase Modulation on Diabetic Kidney Disease at Single-Cell Resolution.” *Cell Reports. Medicine* 4 (4): 100992.

Bankir, L., and C. de Rouffignac. 1985. “Urinary Concentrating Ability: Insights from Comparative Anatomy.” *The American Journal of Physiology* 249 (6 Pt 2): R643–66.

Bittner, Noëlle K. J., Katya L. Mack, and Michael W. Nachman. 2022. “Shared Patterns of Gene Expression and Protein Evolution Associated with Adaptation to Desert Environments in Rodents.” *Genome Biology and Evolution* 14 (11). https://doi.org/10.1093/gbe/evac155.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics* 30 (15): 2114–20.

Bothma J, Matthee S, Matthee C The evolutionary history of parasitic sucking lice and their rodent hosts: A case of evolutionary co‐divergences December 2019 Zoologica Scripta 49(1)

Cornejo-Latorre C,. Cortés-Calva P, and Ticul Álvarez S The evolutionary history of the subgenus Haplomylomys (Cricetidae: Peromyscus) Journal of Mammalogy, Volume 98, Issue 6, 1 December 2017, Pages 1627–1640, https://doi.org/10.1093/jmammal/gyx107

Böttinger, Erwin P. 2010. “Lights on for Aminopeptidases in Cystic Kidney Disease.” *The Journal of Clinical Investigation* 120 (3): 660–63.

Brunskill, Eric W., and S. Steven Potter. 2010. “Gene Expression Programs of Mouse Endothelial Cells in Kidney Development and Disease.” *PloS One* 5 (8): e12034.

Bryja, Josef, Ondřej Mikula, Radim Šumbera, Yonas Meheretu, Tatiana Aghová, Leonid A. Lavrenchenko, Vladimír Mazoch, et al. 2014. “Pan-African Phylogeny of Mus (subgenus Nannomys) Reveals One of the Most Successful Mammal Radiations in Africa.” *BMC Evolutionary Biology* 14 (December):256.

Cao, Junyue, Darren A. Cusanovich, Vijay Ramani, Delasa Aghamirzaie, Hannah A. Pliner, Andrew J. Hill, Riza M. Daza, et al. 2018. “Joint Profiling of Chromatin Accessibility and Gene Expression in Thousands of Single Cells.” *Science* 361 (6409): 1380–85.

Castañeda-Rico, Susette, Jesús E. Maldonado, Melissa T. R. Hawkins, and Cody W. Edwards. 2025. “Unveiling Hidden Diversity: Phylogenomics of Neotomine Rodents and Taxonomic Implications for the Genus Peromyscus.” *Molecular Phylogenetics and Evolution* 203 (February):108233.

Chamberlain, Scott A., and Carl Boettiger. 2017. “R Python, and Ruby Clients for GBIF Species Occurrence Data.” *PeerJ*. https://doi.org/10.7287/peerj.preprints.3304v1.

Chen, Lihe, Jae Wook Lee, Chung-Lin Chou, Anil V. Nair, Maria A. Battistone, Teodor G. Păunescu, Maria Merkulova, et al. 2017. “Transcriptomes of Major Renal Collecting Duct Cell Types in Mouse Identified by Single-Cell RNA-Seq.” *Proceedings of the National Academy of Sciences of the United States of America* 114 (46): E9989–98.

Chevret, Pascale, Terence J. Robinson, Julie Perez, Frédéric Veyrunes, and Janice Britton-Davidian. 2014. “A Phylogeographic Survey of the Pygmy Mouse Mus Minutoides in South Africa: Taxonomic and Karyotypic Inference from Cytochrome B Sequences of Museum Specimens.” *PloS One* 9 (6): e98499.

Chmielewski, Christine. 2003. “Renal Anatomy and Overview of Nephron Function.” *Nephrology Nursing Journal : Journal of the American Nephrology Nurses’ Association* 30 (2): 185–90; quiz 191–92.

Chrysopoulou, Maria, and Markus M. Rinschen. 2024. “Metabolic Rewiring and Communication: An Integrative View of Kidney Proximal Tubule Function.” *Annual Review of Physiology* 86 (February):405–27.

Clark, Jevin Z., Lihe Chen, Chung-Lin Chou, Hyun Jun Jung, Jae Wook Lee, and Mark A. Knepper. 2019. “Representation and Relative Abundance of Cell-Type Selective Markers in Whole-Kidney RNA-Seq Data.” *Kidney International* 95 (4): 787.

Darvish, Jamshid & Mohammadi, Zeinolabedin & Ghorbani, Fatemeh & Mahmoudi, Ahmad & Dubey, Sylvain. (2015). Phylogenetic Relationships of Apodemus Kaup, 1829 (Rodentia: Muridae) Species in the Eastern Mediterranean Inferred from Mitochondrial DNA, with Emphasis on Iranian Species. Journal of Mammalian Evolution. 22. 10.1007/s10914-015-9294-9.

Donald, John, and Thomas L. Pannabecker. 2015. “Osmoregulation in Desert-Adapted Mammals.” *Sodium and Water Homeostasis*, 191–211.

Ewels, Philip A., Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, and Sven Nahnsen. 2020. “The Nf-Core Framework for Community-Curated Bioinformatics Pipelines.” *Nature Biotechnology* 38 (3): 276–78.

Fabre, Pierre-Henri, Lionel Hautier, Dimitar Dimitrov, and Emmanuel J. P. Douzery. 2012. “A Glimpse on the Pattern of Rodent Diversification: A Phylogenetic Approach.” *BMC Evolutionary Biology* 12 (June):88.

Fenton, Robert A., Chung-Lin Chou, Gavin S. Stewart, Craig P. Smith, and Mark A. Knepper. 2004. “Urinary Concentrating Defect in Mice with Selective Deletion of Phloretin-Sensitive Urea Transporters in the Renal Collecting Duct.” *Proceedings of the National Academy of Sciences of the United States of America* 101 (19): 7469–74.

Garba, Madougou, Ambroise Dalecky, Ibrahima Kadaoure, Mamadou Kane, Karmadine Hima, Sophie Veran, Sama Gagare, et al. 2014. “Spatial Segregation between Invasive and Native Commensal Rodents in an Urban Environment: A Case Study in Niamey, Niger.” *PloS One* 9 (11): e110666.

Gillard, Gareth B., Lars Grønvold, Line L. Røsæg, Matilde Mengkrog Holen, Øystein Monsen, Ben F. Koop, Eric B. Rondeau, et al. 2021. “Comparative Regulomics Supports Pervasive Selection on Gene Dosage Following Whole Genome Duplication.” *Genome Biology* 22 (1): 1–18.

Giorello, Facundo M., Matias Feijoo, Guillermo D’Elía, Daniel E. Naya, Lourdes Valdez, Juan C. Opazo, and Enrique P. Lessa. 2018. “An Association between Differential Expression and Genetic Divergence in the Patagonian Olive Mouse (Abrothrix Olivacea).” *Molecular Ecology*, June. https://doi.org/10.1111/mec.14778.

Grabherr, Manfred G., Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, et al. 2011. “Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome.” *Nature Biotechnology* 29 (7): 644–52.

Granjon L., Bâ K., Diagne C., Ndiaye A., Piry S. & Thiam M., 2019. La communauté des petits Rongeurs : tendances historiques et caractéristiques du peuplement actuel. In: La Grande Muraille Verte, une réponse africaine au changement climatique. Boëtsch G, Duboz P, Guissé A, Sarr P (Eds.). CNRS Editions, Paris, 380pp.

Granjon L. & Duplantier J.M., 2009. Les Rongeurs de l’Afrique sahélo-soudanienne. IRD/MNHN (Collection Faune et Flore tropicale n°43), Marseille.

Haas, Brian J., Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D. Blood, Joshua Bowden, Matthew Brian Couger, et al. 2013. “De Novo Transcript Sequence Reconstruction from RNA-Seq Using the Trinity Platform for Reference Generation and Analysis.” *Nature Protocols* 8 (8): 1494–1512.

Habuka, Masato, Linn Fagerberg, Björn M. Hallström, Caroline Kampf, Karolina Edlund, Åsa Sivertsson, Tadashi Yamamoto, Fredrik Pontén, Mathias Uhlén, and Jacob Odeberg. 2014. “The Kidney Transcriptome and Proteome Defined by Transcriptomics and Antibody-Based Profiling.” *PloS One* 9 (12): e116125.

Hafner, John C., Jessica E. Light, David J. Hafner, Mark S. Hafner, Emily Reddington, Duke S. Rogers, and Brett R. Riddle. 2007. “Basal Clades and Molecular Systematics of Heteromyid Rodents.” *Journal of Mammalogy* 88 (5): 1129–45.

Hanna, Ramy M., Rebecca S. Ahdoot, Kamyar Kalantar-Zadeh, Lena Ghobry, and Ira Kurtz. 2021. “Calcium Transport in the Kidney and Disease Processes.” *Frontiers in Endocrinology* 12:762130.

Hao, Yuhan, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck 3rd, Shiwei Zheng, Andrew Butler, Maddie J. Lee, et al. 2021. “Integrated Analysis of Multimodal Single-Cell Data.” *Cell* 184 (13): 3573–87.e29.

Hardouin, Emilie A., Francesca Riccioli, Demetra Andreou, Miguel Baltazar-Soares, Marin Cvitanović, Nathan F. Williams, Pascale Chevret, et al. 2024. “Population Genetics and Demography of the Endemic Mouse Species of Cyprus, Mus Cypriacus.” *Mammalian Biology* 104 (3): 311–22.

He, Y., Hu, S., Ge, D., Yang, Q., Connor, T., & Zhou, C. (2020). Evolutionary history of Spalacidae inferred from fossil occurrences and molecular phylogeny. *Mammal Review*, *50*(1), 11–24. https://doi.org/10.1111/mam.12170

Hijmans, Robert J., Steven Phillips, John Leathwick, and Jane Elith. 2010. “Dismo: Species Distribution Modeling.” *CRAN: Contributed Packages*. The R Foundation. https://doi.org/10.32614/cran.package.dismo.

Huling, Jennifer C., Trairak Pisitkun, Jae H. Song, Ming-Jiun Yu, Jason D. Hoffert, and Mark A. Knepper. 2012. “Gene Expression Databases for Kidney Epithelial Cells.” *American Journal of Physiology. Renal Physiology* 302 (4): F401–7.

Kordonowy, Lauren, and Matthew MacManes. 2017. “Characterizing the Reproductive Transcriptomic Correlates of Acute Dehydration in Males in the Desert-Adapted Rodent, Peromyscus Eremicus.” *BMC Genomics* 18 (1): 473.

Kumar, Sudhir, Michael Suleski, Jack M. Craig, Adrienne E. Kasprowicz, Maxwell Sanderford, Michael Li, Glen Stecher, and S. Blair Hedges. 2022. “TimeTree 5: An Expanded Resource for Species Divergence Times.” *Molecular Biology and Evolution* 39 (8). https://doi.org/10.1093/molbev/msac174.

Lee, Jae Wook, Chung-Lin Chou, and Mark A. Knepper. 2015. “Deep Sequencing in Microdissected Renal Tubules Identifies Nephron Segment-Specific Transcriptomes.” *Journal of the American Society of Nephrology : JASN* 26 (11): 2669–77.

MacManes, Matthew David. 2017. “Severe Acute Dehydration in a Desert Rodent Elicits a Transcriptional Response That Effectively Prevents Kidney Injury.” *American Journal of Physiology. Renal Physiology* 313 (2): F262–72.

Manni, Mosè, Matthew R. Berkeley, Mathieu Seppey, and Evgeny M. Zdobnov. 2021. “BUSCO: Assessing Genomic Data Quality and Beyond.” *Current Protocols* 1 (12): e323.

Marra, Nicholas J., Soo Hyung Eo, Matthew C. Hale, Peter M. Waser, and J. Andrew DeWoody. 2012. “A Priori and a Posteriori Approaches for Finding Genes of Evolutionary Interest in Non-Model Species: Osmoregulatory Genes in the Kidney Transcriptome of the Desert Rodent Dipodomys Spectabilis (banner-Tailed Kangaroo Rat).” *Comparative Biochemistry and Physiology. Part D, Genomics & Proteomics* 7 (4): 328–39.

Marra, Nicholas J., Andrea Romero, and J. Andrew DeWoody. 2014. "Natural Selection and the Genetic Basis of Osmoregulation in Heteromyid Rodents as Revealed by RNA-Seq." *Molecular Ecology* 23 (11): 2699–2711.

Miao, Zhen, Michael S. Balzer, Ziyuan Ma, Hongbo Liu, Junnan Wu, Rojesh Shrestha, Tamas Aranyi, et al. 2021. "Single Cell Regulatory Landscape of the Mouse Kidney Highlights Cellular Differentiation Programs and Disease Targets." *Nature Communications* 12 (1): 2277.

Michaux J, Bellinvia E, Lymberakis P. Taxonomy, evolutionary history and biogeography of the broad-toothed field mouse (*Apodemus*

mystacinus) in the eastern

Mediterranean area based on mitochondrial and nuclear genes, *Biological Journal of the Linnean Society*, Volume 85, Issue 1, May 2005, Pages 53–63, https://doi.org/10.1111/j.1095-8312.2005.00469.x

Mikula, O., Šumbera, R., Aghová, T., Mbau, J.S., Katakweba, A.S., Sabuni, C.A., Bryja, J. (2016). Evolutionary history and species diversity of African pouched mice (Rodentia: Nesomyidae: *Saccostomus*). —*Zoologica Scripta*, 45, 595–617.

Mikula, Ondřej, Violaine Nicolas, Radim Šumbera, Adam Konečný, Christiane Denys, Erik Verheyen, Anna Bryjová, Alan R. Lemmon, Emily Moriarty Lemmon, and Josef Bryja. 2021. "Nuclear Phylogenomics, but Not Mitogenomics, Resolves the Most Successful Late Miocene Radiation of African Mammals (Rodentia: Muridae: Arvicanthini)." *Molecular Phylogenetics and Evolution* 157 (April):107069.

Mikula, O, Bennett, NC, Van Daele, P, Iek, L, Bryja, J, Visser, JH, Jansen van Vuuren, B, umbera, R (2022). Species limits and phylogeographic structure in two genera of solitary African mole-rats Georychus and Heliophobius. *Mol*

Phylogenet Evol, 167:107337.

Pannabecker, Thomas L. 2013. "Comparative Physiology and Architecture Associated with the

Mammalian Urine Concentrating Mechanism: Role of Inner Medullary Water and Urea Transport Pathways in the Rodent Medulla." *American Journal of Physiology. Regulatory, Integrative and Comparative Physiology* 304 (7): R488–503.

———. 2015. "Aquaporins in Desert Rodent Physiology." *The Biological Bulletin* 229 (1): 120–28.

Paradis, Emmanuel, and Klaus Schliep. 2019. "Ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R." *Bioinformatics (Oxford, England)* 35 (3): 526–28.

Park, Jihwan, Rojesh Shrestha, Chengxiang Qiu, Ayano Kondo, Shizheng Huang, Max Werth, Mingyao Li, Jonathan Barasch, and Katalin Suszták. 2018. "Single-Cell Transcriptomics of the Mouse Kidney Reveals Potential Cellular Targets of Kidney Disease." *Science* 360 (6390): 758–63.

Pisano, J., Condamine, F.L., Lebedev, V., Bannikova, A., Quéré, J.-P., Shenbrot, G.I., Pagès, M. and Michaux, J.R. (2015), Out of Himalaya: the impact of past Asian environmental changes on the evolutionary and biogeographical history of Dipodoidea (Rodentia). J. Biogeogr., 42: 856-870. https://doi.org/10.1111/jbi.1 et flores tropicales IRD éditions, 2009 - 215 pages

Revell, Liam J., and Luke J. Harmon. 2022. *Phylogenetic Comparative Methods in R*. Princeton University Press.

Rocha, Joana L., Raquel Godinho, José C. Brito, and Rasmus Nielsen. 2021. "Life in Deserts: The Genetic Basis of Mammalian Desert Adaptation." *Trends in Ecology & Evolution* 36 (7): 637–50.

Rohlfs, Rori V., and Rasmus Nielsen. 2015. "Phylogenetic ANOVA: The Expression Variance and Evolution Model for Quantitative Trait Evolution." *Systematic Biology* 64 (5): 695–708.

Roy, Ankita, Mohammad M. Al-bataineh, and Núria M. Pastor-Soler. 2015. "Collecting Duct Intercalated Cell Function and Regulation." *Clinical Journal of the American Society of*

Nephrology : CJASN 10 (2): 305–24.

Savassi, Boris A. E. S., Gauthier Dobigny, Jonas R. Etougbétché, Thalasse T. Avocegan, François T. Quinsou, Philippe Gauthier, Moudachirou Ibikounlé, Hélène Moné, and Gabriel Mouahid. 2021. “Mastomys Natalensis (Smith, 1834) as a Natural Host for Schistosoma Haematobium (Bilharz, 1852) Weinland, 1858 X Schistosoma Bovis Sonsino, 1876 Introgressive Hybrids.” *Parasitology Research* 120 (5): 1755–70.

Schlöndorff, Detlef, and Bernhard Banas. 2009. “The Mesangial Cell Revisited: No Cell Is an Island.” *Journal of the American Society of Nephrology : JASN* 20 (6): 1179–87.

Sikes, Robert S., and Animal Care and Use Committee of the American Society of Mammalogists. 2016. “2016 Guidelines of the American Society of Mammalogists for the Use of Wild Mammals in Research and Education.” *Journal of Mammalogy* 97 (3): 663–88.

Sohara, Eisei, Tatemitsu Rai, Sei Sasaki, and Shinichi Uchida. 2006. “Physiological Roles of AQP7 in the Kidney: Lessons from AQP7 Knockout Mice.” *Biochimica et Biophysica Acta* 1758 (8): 1106–10.

Stefańska, Ania, Bruno Péault, and John J. Mullins. 2013. “Renal Pericytes: Multifunctional Cells of the Kidneys.” *Pflugers Archiv : European Journal of Physiology* 465 (6): 767–73.

Strutz, Frank, and Michael Zeisberg. 2006. “Renal Fibroblasts and Myofibroblasts in Chronic Kidney Disease.” *Journal of the American Society of Nephrology : JASN* 17 (11): 2992–98.

Swanson, Mark T., Carl H. Oliveros, and Jacob A. Esselstyn. 2019. “A Phylogenomic Rodent Tree Reveals the Repeated Evolution of Masseter Architectures.” *Proceedings. Biological Sciences* 286 (1902): 20190672.

Thiagarajan, Rathi D., Kylie M. Georgas, Bree A. Rumballe, Emmanuelle Lesieur, Han Sheng Chiu, Darrin Taylor, Dave T. P. Tang, Sean M. Grimmond, and Melissa H. Little. 2011. “Identification of Anchor Genes during Kidney Development Defines Ontological Relationships, Molecular Subcompartments and Regulatory Pathways.” *PloS One* 6 (2): e17286.

Tigano, Anna, Jocelyn P. Colella, and Matthew D. MacManes. n.d. “Comparative and Population Genomics Approaches Reveal the Basis of Adaptation to Deserts in a Small Rodent.” https://doi.org/10.1101/856310.

Upham, N.S. & Patterson, B.D. (2015) Evolution of caviomorph rodents: a complete phylogeny and timetree for living genera. In: Vassallo, A.I. and Antenucci, D. (Eds.) Biology of Caviomorph Rodents: Diversity and Evolution. Buenos Aires, SAREM Series A, 63–120.

Upham, Nathan S., Jacob A. Esselstyn, and Walter Jetz. 2019. “Inferring the Mammal Tree: Species-Level Sets of Phylogenies for Questions in Ecology, Evolution, and Conservation.” *PLOS Biology* 17 (12): e3000494.

Urity, Vinoo B., Tadeh Issaian, Eldon J. Braun, William H. Dantzler, and Thomas L. Pannabecker. 2012. “Architecture of Kangaroo Rat Inner Medulla: Segmentation of Descending Thin Limb of Henle’s Loop.” *American Journal of Physiology. Regulatory, Integrative and Comparative Physiology* 302 (6): R720–26.

Vezzoli, Giuseppe, Annalisa Terranegra, Francesco Rainone, Teresa Arcidiacono, Mario Cozzolino, Andrea Aloia, Elena Dogliotti, Daniele Cusi, and Laura Soldati. 2011. “Calcium-Sensing Receptor and Calcium Kidney Stones.” *Journal of Translational Medicine* 9 (November):201.

Wang, S. K., B. M. Reid, S. L. Dugan, J. A. Roggenbuck, L. Read, P. Aref, A. P. H. Taheri, M. Z. Yeganeh, J. P. Simmer, and J. C-C Hu. 2014. “FAM20A Mutations Associated with Enamel Renal Syndrome.” *Journal of Dental Research* 93 (1): 42–48.

Wang, Xuran, Jihwan Park, Katalin Susztak, Nancy R. Zhang, and Mingyao Li. 2019. “Bulk Tissue Cell Type Deconvolution with Multi-Subject Single-Cell Expression Reference.” *Nature Communications* 10 (1): 380.

Wu, Tianzhi, Erqiang Hu, Shuangbin Xu, Meijun Chen, Pingfan Guo, Zehan Dai, Tingze Feng, et al. 2021. “clusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data.” *The Innovation*. https://doi.org/10.1016/j.xinn.2021.100141.

Zelditch, Miriam L., Jingchun Li, Lucy A. P. Tran, and Donald L. Swiderski. 2015. “Relationships of Diversity, Disparity, and Their Evolutionary Rates in Squirrels (Sciuridae).” *Evolution; International Journal of Organic Evolution* 69 (5): 1284–1300.

Zhou, Le-Ting, Shen Qiu, Lin-Li Lv, Zuo-Lin Li, Hong Liu, Ri-Ning Tang, Kun-Ling Ma, and Bi-Cheng Liu. 2018. “Integrative Bioinformatics Analysis Provides Insight into the Molecular Mechanisms of Chronic Kidney Disease.” *Kidney & Blood Pressure Research* 43 (2): 568–81.